

AUDIOVISUAL SPEECH PROCESSING: IMPLICATIONS FOR SPEECH PERCEPTION AND LANGUAGE DEVELOPMENT

**A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville**

**Ryan A. Cannistraci
August 2021**

Copyright © 2021 by Ryan A. Cannistraci
All rights reserved.

DEDICATION

This dissertation is dedicated to my grandparents, John and Camille Sweeney and John and Joanne Cannistraci. Thank you for instilling in me the importance of education and hard work, I hope this manuscript makes you proud.

ACKNOWLEDGEMENTS

Throughout my graduate education and the writing of this dissertation, I have been so fortunate to receive mentorship and support from a vast network of phenomenal people and scientists. I would like to first thank all the families who participated in the experiments included in this manuscript. Without your help, this work would not have been possible.

I would like to thank Dr. Daniela Corbetta, Dr. Todd Freeberg, and Dr. Caglar Tas for their mentorship, collaboration, and letter-writing.

I am especially grateful for the members of my comprehensive exam and dissertation committees, Dr. Shannon Ross-Sheehy, Dr. Greg Reynolds, and Dr. Devin Casenhiser—your interest in my dissertation project and career development has been invaluable to me.

Next, I want to thank the past and present members of the Infant Language and Perceptual Learning Lab. I am especially thankful for my lab mates Ferhat, Sara, Rodrigo, Becca, and Kenzie for their warmth, friendship, and support over the past 7 years. I am similarly thankful for our lab managers, Dora, Rebecca, and Daniela for their constant research support, but also the chats and coffee breaks—thank you all for keeping me relatively sane.

Dr. Aaron Buss, thank you for taking the time to mentor me and regularly meet with me over the past 7 years to develop my research interests as well as modeling and neuroimaging technical skills. I am so thankful for your mentorship.

Dr. Jessica Hay, thank you from the bottom of my heart for accepting me to work in your lab and for mentoring me through my graduate education. I have had such a great experience learning from you—you are an excellent and caring scientist and an even better person. I can only hope to provide my future students with the same discerning and enthusiastic mentorship.

Finally, I would like to thank my family and friends for their love and support. To my friends (Keri, Bret, Jason, and many more), thank you for your friendship and for being my sounding board. To my siblings Dan, Stephen, and Lauren, thank you for always having my back and not letting me take myself too seriously. Mom, thank you for inspiring me to chase my dreams, and Dad, thank you for *a/ways* encouraging me and believing in me. Brittany, my dear, thank you for loving me and caring for me physically and emotionally through the final throes of this project, I would not have been able to do this without you. I am so grateful that we have taken this adventure together and I can't wait to see what's next.

ABSTRACT

This dissertation aims to empirically assess the complex, multileveled relationships between audiovisual speech perception and early language development. The majority of extant language development research has justifiably focused on infants' ability to learn language from auditory input, and indeed, infants are precocious auditory learners (Saffran & Kirkham, 2018). Complementary to auditory speech, however, are the necessarily redundant facial movements used to articulate speech. Outside of language development research, multimodal processing has been theorized to facilitate perceptual learning and cognitive development (Bahrick & Lickliter, 2000), but only a small number of empirical studies have investigated how audiovisual speech perception in infancy is related to later language development. Infants demonstrate sensitivity to audiovisual speech early in development (Patterson & Werker, 2003), however, it is still an open question as to how infants may functionally use these redundancies in the service of language learning. This dissertation first examines how infants integrate auditory and visual speech at 4.5-months by measuring their perception of the McGurk effect. To augment this behavioral work, a meta-analysis assesses infants' perception of the McGurk Effect across available published and unpublished literature. A dynamic neural field (DNF) computational model then explores the real-time perceptual and neuro-cognitive processes that lead to the perception of the McGurk Effect. Next, to explore how audiovisual speech perception in infants relates to later language outcomes, longitudinal relationships between infants' perception of the McGurk Effect, their native and non-native phoneme discrimination at 7-months, and vocabulary outcomes at 13- and 18-months are explored. Finally, the utility of audiovisual redundancies was tested for infants in a difficult object-label association laboratory task. The culmination of this dissertation creates a body of empirical and computational work focused on infants' audiovisual speech perception at multiple timepoints across children's burgeoning language ability in the first two years of life.

TABLE OF CONTENTS

INTRODUCTION	1
CHAPTER I Infants' Individual Differences in Perceiving the McGurk Effect	6
CHAPTER II A Meta-Analytic Investigation of Infants' Perceptions of the McGurk Effect.....	15
CHAPTER III A Dynamic Neural Field Exploration of Individual Differences in the McGurk Effect	27
CHAPTER IV Longitudinal Correspondances Between Infants' Audiovisual Speech Percpetion and Native Language Perceptual Narrowing	37
CHAPTER V 14-Month-Olds' Learning of Minimal Pair Object-Label Associations from Audiovisual Speech	53
CONCLUSION	66
Chapter 1	66
Chapter 2	66
Chapter 3	67
Chapter 4	67
Chapter 5	68
Overall Conclusions and Future Directions	68
References	71
Appendix.....	79
VITA.....	82

LIST OF TABLES

Table 1. Model summary descriptions of all growth mixture models, with the best fitting model (m6) in italics.	48
Table 2. Model summary of the best fitting growth mixture model (m6).	49
Table 3. Bayesian ANCOVA model comparison.	64

LIST OF FIGURES

<i>Figure 1.1.</i> Trial layout for the McGurk Effect infant habituation study.	11
<i>Figure 1.2.</i> Infants' looking time difference scores for /ba/ and /da/ test trials to demonstrate individual differences in infants' McGurk Effect perception, clustered by looking behavior.	12
<i>Figure 2.1.</i> Forest plot depicting the average meta-analytic effect of the random effect model (SMD), as well as the effect size (TE), standard error (seTE), confidence interval (95%-CI), and model weighting (Weight) of each included research record.	19
<i>Figure 2.2.</i> Funnel plot that graphs each record in the meta-analysis by effect size (Hedge's g) and Standard Error.	20
<i>Figure 2.3.</i> P-Curve plot including the statistically significant records from in the meta-analysis. These statistically significant records pass right-skewness and flatness tests, suggesting that the significant results are distributed in a pattern that does not suggest evidence of data manipulation and that there is sufficient power to detect these significant results.	21
<i>Figure 2.4.</i> A forest plot for stimulus combination subgroup analysis with fusion stimuli at the top, audiovisual mismatch stimuli in the middle, and visually-dominant stimuli at the bottom. SMD refers to the mean effect size for each of the subgroups and the prediction interval spans the range where effect sizes using these stimuli combination would be expected to be normally distributed.	22
<i>Figure 2.5.</i> Forest plot for experimental design subgroup analysis, including EEG methods at the top, Habituation methods in the middle, and Intermodal looking methods at the bottom. SMD refers to the mean effect size for each subgroup and the prediction interval is the expected range for normally distributed effect sizes for future work using the three experimental methods.	24
<i>Figure 2.6.</i> Forest plot for between (top) vs within-subjects (bottom) subgroup analysis. SMD is the mean effect size for each subgroup and the prediction interval spans the expected range wherein effect sizes may be normally distributed.	25
<i>Figure 3.1.</i> The DNF model architecture of the infant McGurk Effect model at resting state. Hebbian learning pre-shapes for phonemes across each field are labeled.	30
<i>Figure 3.2.</i> The DNF model architecture of the infant McGurk Effect model binding the fusion effect in the Perceptual Field, positive activation in the Attention Node, and a working memory peak forming below threshold in the Working Memory Field.	31
<i>Figure 3.3.</i> Adult McGurk Effect model simulation results compared to observed rates in the extant literature.	34

<i>Figure 4.1.</i> Elbow method plot for k-means cluster values derived from the lowest Bayesian Information Criterion value which suggests that four clusters are optimal in the observed data.....	44
<i>Figure 4.2.</i> Elbow method plot for k-means cluster values derived from the lowest Akaike Information Criterion value which also suggests that four clusters are optimal in the observed data.....	45
<i>Figure 4.3.</i> Mean and standard deviation values for the auditory and fusion preference data that compose each of the four clusters.	46
<i>Figure 4.4.</i> A t-Distributed Stochastic Neighbor Embedding (t-SNE) plot of the four clusters defined from our longitudinal data. Note: t-SNE reduces high-dimensional data to a 2D visual representation, thus, the specific organization shown here is conceptual.	47
<i>Figure 4.5.</i> Clustered mean growth model Z-Scores indexed by Measure across the first 18 months of life. Note: Dark grey ranges indicate 95% confidence intervals.	50
<i>Figure 5.1.</i> Bar graph depicting infants' looking time for Same and Switch trials across experimental conditions.....	61
<i>Figure 5.2.</i> Scatterplot depicting the correlation between infants' time to habituate and their Same-Switch difference scores.	63

INTRODUCTION

Human infants are prodigious learners across many domains. In the study of language development, historical and contemporary approaches have focused on how infants learn language from auditory signals in their surrounding environment. These approaches have been especially fruitful to uncover how infants home in on the relevant sounds of their native language(s) (e.g., Maye, Werker, & Gerken, 2002), segment words from continuous speech streams (e.g., Saffran, Aslin, & Newport, 1996), and map auditory labels to objects (e.g., Stager & Werker, 1997). The developmental literature often characterizes the infants' language learning environment as "one great blooming, buzzing confusion" (James, 1890). Despite this sensory confusion, infants track regularities in the language they hear amongst this bustling environment and learn language with relative ease. The acoustic information available in an infants' perceptual environment is not the only structured source of information that can be used for learning. In fact, by taking the broader context around James's (1890) original quote, it becomes apparent that the sensory confusion around the developing infant contains structured, multimodal information.

"The baby, assailed by eyes, ears, nose, skin, and entrails at once, feels it all as one great blooming, buzzing confusion; and to the very end of life, our location of all things in one space is due to the fact that the original extents or bignesses of all the sensations which came to our notice at once, coalesced together into one and the same space. There is no other reason than this why 'the hand I touch and see coincides spatially with the hand I immediately feel'" (James, 1890).

The full context of James's (1890) quote suggests that an infant's sensory environment is fundamentally multisensory, and that temporally synced, multimodal information plays an integral role in perceptual learning. A similar thesis and multimodal focus for perception and cognition is especially pertinent to developmental research, since infants are born into a multisensory world. Gibson's ecological approach to development (e.g., Gibson, 1988) can be viewed as a precursor to Bahrick and Lickliter's (2000) *Intersensory Redundancy Hypothesis (IRH)*. Gibson (2000, 2003) posited that multimodal information is both a useful and an economical guiding force for infants' learning. A redundant stimulus from more than one sense modality specifies relevant information for learning across development (Gibson, 2000). In the *Intersensory Redundancy Hypothesis (IRH)*, Bahrick and Lickliter (2000) postulate that infants can use redundant sensory information to learn about amodal properties of stimuli (i.e., information not specific to a single sensory modality). Empirical support for the *IRH* comes from 3-month-olds' learning of rhythmic patterns from audiovisual stimuli (Bahrick, Flom, & Lickliter, 2002). Bahrick and colleagues (2002) first habituated infants to either an audiovisual event or a visual-only event where the amodal property of the stimuli was the tempo at which a hammer tapped a

surface. Infants in the audiovisual (i.e., bimodal) condition received auditory information as the hammer reached the bottom of its trajectory—the sound of a hammer hitting a wooden surface—in addition to the visible motion of the hammer. Infants in the visual-only (i.e., unimodal) condition simply saw the hammer’s motion. Infants were able to detect change in tempo (i.e., the amodal information) when habituated in the audiovisual condition but were unsuccessful in the visual-only condition. The redundant, audiovisual sensory information facilitated perceptual learning. By five months of age, infants were able to detect tempo changes when habituated to both audiovisual and visual-only stimuli in the same task (Bahrick & Lickliter, 2002, 2004). Cumulatively and consistent with the *IRH*, these findings demonstrate that younger infants preferentially attend to bimodal information and learn about the amodal properties of the stimuli, while older infants can flexibly attend to both bimodal and unimodal cues due to their accumulated perceptual experience through sensory-motor activity across the first year of life (Bahrick & Lickliter, 2004). Interestingly, in difficult learning scenarios, older infants who typically learn from unimodal cues revert to preferring bimodal information, since the redundant multisensory information better supports learning (Bahrick, Lickliter, Castellanos, & Vaillant-Molina, 2010).

Human speech contains “inherently multimodal” (Rosenblum, 2008) audiovisual redundancies—the facial movements that produce speech as well as the speech sounds themselves coalesce in both time and space. The auditory and visual speech signals convey redundant information about the intended speech percept. This audiovisual redundancy is useful for enhancing speech perception for adults when in noisy, challenging auditory environments (Sumby & Pollack, 1954). The seminal work from Bahrick and colleagues on the *IRH* (Bahrick & Lickliter, 2000, 2002, 2004) suggest infants leverage ecologically-relevant, multimodal information during learning, and it may in fact be the case that infants leverage audiovisual speech information to facilitate learning language. Indeed, Bahrick, Todd, and Soska (2018) found that infants with greater average looking duration and more accurate multimodal matching accuracy had larger receptive and productive vocabulary sizes between the ages of two and five, measured by the Mullen Scales of Early Learning (Mullen, 1995). These findings suggest that quicker, more accurate multimodal matching may be relevant for language development outcomes.

Audiovisual Speech in Infancy

Infants display a robust sensitivity to the audiovisual redundancies in speech early in the first year of life. At two months, infants readily link vowel sounds to matching visual facial movements by preferentially attending to a correctly articulating face as opposed to an asynchronous distractor (Patterson & Werker, 2003). Four-month-olds detect audiovisual asynchrony in speech perception (Lewkowicz, 2010) and 5-month-old infants preferentially attend to congruent audiovisual continuous speech (Kuhl & Meltzoff, 1984). At six-months, visual articulatory information enhances phoneme discrimination, suggesting that

audiovisual redundancies may augment the learning of phonetic boundaries in infancy (Teinonen, Aslin, Alku, & Csibra, 2008). In addition to being sensitive to audiovisual speech, infants can use visual information alone to discriminate their native language from an unknown language (Weikum, Vouloumanos, Navarra, Soto-Faraco, Sebastián-Gallés, & Werker, 2007). In a study by Weikum and colleagues (2007), 4-, 6-, and 8-month-old infants were habituated to visual utterances in a single language (e.g., English) and were tested on their ability to distinguish an utterance in the same language as habituation from a second utterance in a different language (e.g., French). Both 4- and 6-month-olds looked for a longer duration to the switch trial (foreign language) than the same trial (native language), suggesting the younger infants in this study were able to discriminate between their native language and a foreign language based on visual information alone. However, the 8-month-olds did not show evidence of discriminating their native language from visual speech. In a follow-up experiment, bilingual and monolingual 8-month-olds were tested in the same experimental design and the bilingual 8-month-olds were able to discriminate their native from a foreign language using visual speech information, while the monolingual infants were not. Together, these results suggest that infants may begin categorizing the visual speech information of their native language in the first year of life, giving rise to the pattern of findings that younger infants can discriminate languages using visual information, but this sensitivity diminishes around 8 months. Perceptual narrowing, where infants home in on meaningful perceptual information in their environment and learn to ignore perceptual distinctions that are not relevant (Maurer & Werker, 2014), is hypothesized to be an important driver of learning across many domains in infancy. Bilingual infants often show a pattern of delayed perceptual narrowing in many domains (for a review, see Byers-Heinlein & Fennell, 2014), and the monolingual vs bilingual results from Weikum and colleagues (Weikum et al., 2007) suggest a similar narrowing process might occur for sensitivity to visual speech.

While infants exhibit sensitivity to audiovisual speech congruencies in laboratory tasks, some researchers (e.g., Shaw & Bortfeld, 2015; Desjardins & Werker, 2004) argue that these empirical findings do not necessarily indicate that infants integrate auditory and visual speech information. Sensitivity to audiovisual speech information is necessary, but not sufficient evidence to suggest that infants integrate auditory and visual speech information to their phoneme or lexical representations in a meaningful manner. However, infants' sensitivity to a canonical audiovisual speech illusion, the McGurk Effect (McGurk & MacDonald, 1976) does suggest that infants readily integrate auditory and visual speech information during speech perception (e.g., Burnham & Dodd, 1996, 2004; Rosenblum, Schmuckler, & Johnson, 1997). In the classic example of the McGurk Effect, McGurk & MacDonald (1976) inadvertently violated the typically redundant nature of audiovisual speech information while dubbing audio and video recordings, which created illusory percepts of audiovisual speech. The original McGurk effect was composed of the syllables /ba/ and /ga/. These two

syllables are produced with different places of articulation and elicit distinctive facial movements—/ba/ is bilabial, meaning the consonant closure occurs at the speaker's lips and /ga/ is velar, meaning the consonant closure occurs as the back of the tongue contacts the velum (top of the mouth). In McGurk and MacDonald's (1976) seminal study, pre-school children and adult participants reported experiencing the syllable /da/ when viewing an audiovisual stimulus composed of an auditory /ba/ and visual /ga/. Interestingly, the emergent /da/ percept was not actually present in either the auditory or visual signal and was thus considered a fused or illusory percept. Since McGurk and MacDonald's (1976) original study, there have been several studies that generalize the McGurk effect to infant populations using different stimuli combinations and experimental paradigms (e.g., Burnham & Dodd, 1996, 2004; Desjardins & Werker, 2004; Rosenblum, Schmuckler, & Johnson, 1997).

There is ample evidence to suggest that infants are sensitive to the audiovisual redundancies of speech and that infants integrate auditory and visual information in McGurk Effect contexts. However less is known about how audiovisual speech perception may be functionally related to language development, although several researchers have suggested such a connection. For example, Teinonen and colleagues (Teinonen, Aslin, Alku, & Csibra, 2008) demonstrated that access to visual articulatory information enhanced 6-month-olds' ability to form distinct phoneme categories. From these findings, Teinonen and colleagues (2008) hypothesized that redundant visual speech information may be facilitative for forming phoneme categories. As an infant accumulates language experience, endogenous factors may also influence preferences for audiovisual speech. Lewkowicz and Hansen-Tift (2012) have recently described a trend of infants' looking preference moving from a speaker's eyes to their mouth around 8 months. Lewkowicz and Hansen-Tift (2012) hypothesized that the timing of this shift is driven by the onset of canonical babbling and perceptual narrowing of native language phonemes—two early language learning processes that may be facilitated by redundant audiovisual speech information. Indeed, correlational evidence suggests that infants who preferentially look at a speaker's mouth at 5-6 months have larger expressive vocabularies at 24 months (Young, Merin, Rogers, & Ozonoff, 2009) and at 5 years of age (Golan et al., 2018). These findings suggest that audiovisual speech perception in infancy is empirically interesting because individual differences in audiovisual speech perception may be related to later language outcomes.

Current Research

This dissertation employs behavioral and computational methods to explore the relationship between audiovisual speech perception and language learning across the first two years of life. Chapter 1 utilizes a within-subjects design to measure audiovisual speech integration for 4.5-month-old infants' perception of the McGurk Effect. Chapter 2 further explores the literature on infants' perception of the McGurk Effect using a meta-analysis to quantify the

average expected effect size, experimental task differences, and publication biases in the extant literature. Chapter 3 examines potential individual differences in McGurk Effect perception using a *Dynamic Neural Field (DNF)* computational model. The DNF model sheds light on the relevant individual differences by manipulating neurocomputational parameters that contribute to audiovisual speech perception. Chapter 4 explores longitudinal correspondences between infants' perception of the McGurk Effect at 4.5 months, their native and non-native audiovisual phoneme discrimination at 7 months, as well as vocabulary size measures at 13 and 18 months using planned correlational analyses and statistical growth mixture modeling. These analyses aim to illuminate how early audiovisual speech perception might predict later language outcomes. Finally, in Chapter 5, a new cohort of 14-month-old infants participated in an audiovisual minimal pair object-label associative learning task that explicitly tests the task difficulty prediction of the IRH (Bahrick, Lickliter, Castellanos, & Vaillant-Molina, 2010). Taken together, the body of work undertaken in this dissertation informs a synthesis between contemporary theories of speech perception and language development, and more importantly, the interwoven relationships between perceptual learning and cognitive development as infants learn their native language(s).

CHAPTER I
INFANTS' INDIVIDUAL DIFFERENCES IN PERCEIVING THE
MCGURK EFFECT

The extant literature on infants' perception of the McGurk Effect provides evidence that infants between 4 and 5 months of age readily perceive the canonical fusion McGurk Effect (Burnham & Dodd, 1996, 2004) as well as visually-dominant illusory percepts (Desjardins & Werker, 2004; Rosenblum, Schmuckler, & Johnson, 1997). Though the McGurk illusion is synthetically induced, its presence may be telling of how humans represent audiovisual speech information (see also Alsius et al., 2017 for a tempered view and further discussion of the applications of the McGurk Effect). Seminal findings in the infant McGurk Effect literature use habituation paradigms to test infants' perception of the McGurk Effect. Infants are often habituated to the McGurk illusion and after habituation, given audiovisually congruent test trials (e.g., /ba/, /da/, or /tha/ is in Burnham & Dodd, 1996, 2004). If infants treat /da/ or /tha/ test trials as familiar (short looking time), and the /ba/ test trial as novel (longer looking time), this pattern of looking behavior suggests that infants perceived the canonical fused /da/ during the habituation trials. More recently, researchers have utilized intermodal looking paradigms using eye-tracking to examine whether infants prefer congruent audiovisual speech, or incongruent audiovisual speech that elicits the McGurk Effect (e.g., D'Souza et al., 2016; Mercure et al., 2019; Tomalski et al. 2013) as well as event-related neural measures such as electroencephalograms (EEG: e.g., Kushnerenko et al., 2008, Kushnerenko et al., 2008) or near-infrared spectroscopy (NIRS: e.g., Ujiie, Kanazawa, & Yamaguchi, 2020).

While the empirical data across studies suggests that infants readily perceive the McGurk Effect, these studies are limited in their theoretical interpretation for why infants perceive this audiovisual illusion. Some studies conclude that infants perceive the McGurk effect through nativistic, articulatory-based representations that need not be learned through perceptual experience (e.g., Rosenblum et al., 1997). Articulatory-based interpretations, however, are unable to explain why some adults do not perceive the McGurk Effect (approximately 16% of adult participants as calculated by Alsius et al., 2017). Bayesian-inference models of speech perception such as Massaro (1987) and, more recently, Magnotti and Beauchamp (2017) require lexical knowledge as a pre-requisite for perceiving the McGurk Effect. These Bayesian-inference accounts aim to explain the McGurk Effect in adult participants, but this interpretation is problematic when considering the developmental data. It is speculative that 4- to 5-month-olds have robust lexical representations from which to base their phoneme perception. The earliest evidence of lexical knowledge in infants is based on findings that 6 to 9-month-olds accurately direct their gaze to pictures of common objects when the corresponding labels are spoken by a caregiver (Bergelson & Swingley, 2012). Alternatively, Burnham and Dodd (1996; 2004) offer the interpretation that audiovisual integration occurs early in sensory processing, before phonological and lexical processing. Burnham and Dodd's (1996; 2004) account offers an alternative to articulatory-based speech perception that can still account for infants' ability to perceive the

McGurk Effect. The overarching research goal of Chapter 1 is to measure infants' audiovisual integration ability for speech stimuli, as evidenced by their susceptibility to the McGurk Effect. These data will be helpful in future discussions of developmental theories of speech perception, especially as these theories relate to ecologically relevant, multimodal information.

In the current study, I tested the rate at which 4.5-month-old infants perceive the fusion effect from McGurk Effect stimuli. I hypothesized that if an infant perceives the fusion McGurk Effect during habituation, they should treat the non-alternating fusion test trial (i.e., habituation stimulus) and the alternating fusion-congruent /da/ test trial as familiar, evidenced by short looking times during testing (i.e., failure to dishabituate). If the infant perceives the McGurk fusion effect, they should restore their looking time (i.e., dishabituate) to the alternating test trial that contains the congruent /ba/, since /ba/ should be a novel percept. These behavioral results will be expressed as difference scores, calculated by subtracting the looking time for the non-alternating fusion test trial in each block from the looking time for the corresponding alternating congruent trial (/da/ or /ba/) within that block. Thus, each infant will have two difference score measures. I predicted that infants who fail to fuse the auditory /ba/ and visual /ga/ will show the opposite pattern of responses—they should dishabituate on alternating trials containing the congruent /da/, as this will be perceived as a novel stimulus, but fail to dishabituate on trials that contain the congruent /ba/. This pattern of behavior would be consistent with adult findings that non-fusers typically perceive the auditory signal (i.e., /ba/) when viewing McGurk Effect stimuli (Alsius et al., 2017).

Methods

Participants

Participants were recruited through the Child Development Research Group (CDRG) database housed and curated by CDRG at The University of Tennessee, Knoxville. In total, 43 monolingual English learning infants between 4 & 5-months of age ($M=4.5$ months) were recruited. Of the infants who participated in Chapter 1, 33 infants (16 girls, 17 boys) were included in the final data set. The additional 10 infants were excluded from the final sample due to excessive fussiness (7) or failure to habituate (3). All participants were monolingual English-learners with fewer than 2 ear infections and no known congenital hearing or vision complications.

An *a priori* power analysis based on a previous work testing the McGurk effect in infants (Rosenblum, Schmuckler, & Johnson, 1997) suggests that $n=16$ is sufficient for observed statistical power of 0.80. However, since Chapter 1 utilizes a different experimental paradigm than Rosenblum and colleagues (1997), a second power analysis for the same paradigm used in Chapter 1 (Tyler, Best, Goldstein, & Antoniou, 2014) suggests $n=20$ is adequate for *a priori* statistical power. Since the data from Chapter 1 will be used in later chapters of this dissertation for correlation analyses and longitudinal growth mixture

modeling, I purposely over-sampled beyond 20 participants to gain statistical power for all planned analyses in future chapters of this dissertation and to compensate for attrition in our longitudinal data.

Stimuli

All stimuli were presented audiovisually and consisted of a speaker's face articulating a syllable. Auditory stimuli were recorded using a Blue Snowball microphone, then cleaned and normalized in Praat (Boersma & Weenink, 1996). Visual stimuli were recorded using a Cannon Vixia HFM40 camera. Auditory and visual recordings were digitally synced together to create the audiovisual stimuli in Adobe Premiere. There were two types of trials, audiovisually congruent trials and McGurk Effect fusion trials. On the audiovisually congruent trials, there were two novel stimuli with matching auditory and visual information speech information (i.e., audiovisual /ba/ and audiovisual /da/). On McGurk Effect trials, the auditory and visual information was incongruent (i.e., auditory /ba/, visual /ga/, typically perceived as /da/). These three stimuli were chosen to differentiate if infants perceive an auditory /ba/ percept or the fusion /da/ percept during McGurk Effect habituation trials.

Apparatus

The experimental sessions were housed in a dimly-lit, sound-controlled room and all stimuli were presented on a high-definition monitor and through high-quality speakers. The experimental procedure was run using Habit2 (Oakes, Sperka, DeBolt, & Cantrell, 2019) and infants' looking time for each trial was coded online in the Habit2 software by the experimenter.

Procedure

Infants were seated on a caregiver's lap throughout the experimental session where they viewed and heard the experimental stimuli. Infants were habituated and tested using an infant-controlled, modified stimulus alternating paradigm (Best & Jones, 1998; Tyler, Best, Goldstein, & Antoniou, 2014). The stimulus alternating paradigm was chosen for this experiment because it was designed to test two stimulus contrasts in a within-subjects design. This within-subjects design yields greater statistical power than previously conducted research on infants' perception of the McGurk effect that used between-subjects designs (e.g., Desjardins & Werker, 2004; Rosenblum, Schmuckler, & Johnson, 1997). Each 4.5-month-old was habituated to the canonical McGurk Effect (auditory /ba/, visual /ga/), and for the habituation phase, I used the same infant-controlled habituation criterion as Singh and colleagues (Singh, Fu, Tay, & Golinkoff, 2018). Habituation criterion was reached when each infant's look duration, averaged across a 3-trial sliding window, decreased by 50% from the average of the first 3 habituation trials. Infants had a 20-trial maximum to reach the habituation criteria. Each trial in this design was infant-controlled, meaning the infants' looking behavior determined the duration of each trial. Trials were

terminated when the infant looked away from the stimuli for over 1 second or if the trial reached its maximum run-time, which was 21.5 seconds. The maximum trial-length encompassed 10 unique presentations of the audiovisual stimuli.

Following the habituation phase, each infant continued through three additional phases composed of two test blocks with an interspersed re-familiarization block—test blocks and trial order were pseudorandomized across participants (*Figure 1.1*). Each test block consisted of two trials, a non-alternating McGurk fusion trial (i.e., same as the habituation stimulus) and an alternating trial where the original fusion stimulus was alternated with a novel test stimulus. In one test block the alternating trial alternated between the fusion stimulus and an AV congruent /da/ stimulus. In the other test block, the alternating trial alternated between the fusion stimulus and an AV congruent /ba/ stimulus. Between each test block, the re-familiarization phase was composed of three trials of the fusion stimulus—the same stimulus from habituation. Infants could potentially see and hear up five unique tokens of each of the fusion and the congruent (/da/ or /ba/) stimuli on alternating test trials. Each infant's looking time data was coded online during the experimental sessions by the author of this dissertation, and his button-press responses to measure each infant's looking times were recorded in Habit2 (Oakes, Sperka, DeBolt, & Cantrell, 2019). The experimenter's button-press and release responses for when the infants were attending to the experimental stimuli or looking away allows for comparisons of each infant's looking time for all trials in the habituation and testing phases.

Results

To assess infants' perception of the McGurk Effect, I first conducted a 2x2 repeated-measures ANOVA (stimulus order by stimulus contrast). There were no significant main effects of stimulus order ($F(1,31)=0.010$, $p=0.922$, $\eta^2<0.001$) or stimulus contrast ($F(1,31)=1.745$, $p=0.196$, $\eta^2=0.029$), nor was there a significant interaction ($F(1,31)=0.773$, $p=0.386$, $\eta^2=0.013$), so I collapsed these data across conditions. I then used follow-up, single samples t-tests to measure looking time differences between the alternating /ba/ test trial and the alternating /da/ test trials compared to zero, as previously used for this experimental design (e.g., Tyler, Best, Goldstein, & Antoniou, 2014) since infants could ostensibly dishabituate to both the /ba/ and /da/ test stimuli. Follow-up single sample t-tests revealed that infants' difference scores for alternating /da/ test trials were significantly greater than zero, $t(32)=3.253$, $p=0.003$, $d=0.566$, but their difference scores for /ba/ test trials were not different from zero, $t(32)=1.190$, $p=0.243$, $d=0.207$. These results suggest that infants significantly restored their looking time to /da/ test trials, meaning that at a group level, infants perceived the McGurk stimuli during habituation as the auditory stimulus /ba/. Interestingly, infants' individual data suggests there may be individual differences in the infants' perception for the McGurk Effect (*Figure 1.2*). I classified infants who responded as if /ba/ was novel as fusers ($n=11$) and infants who responded as if

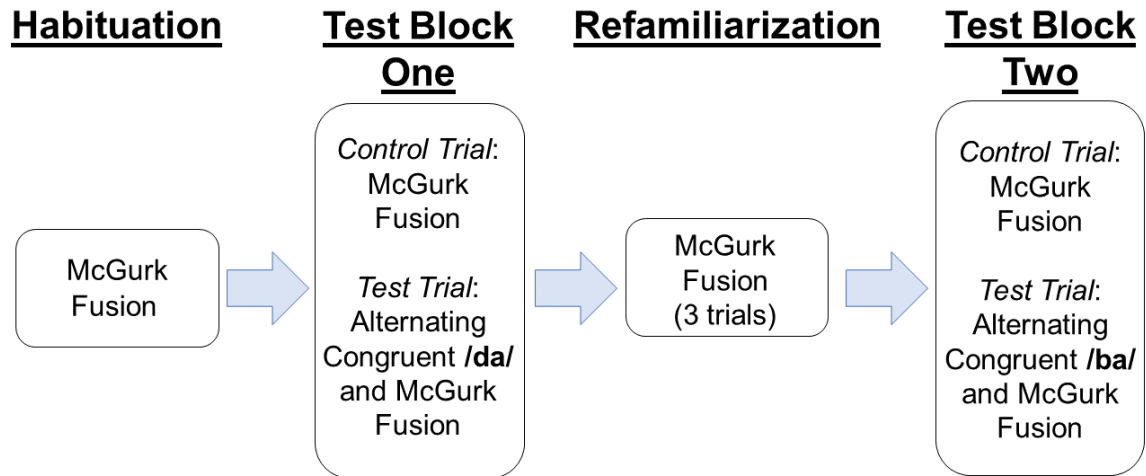


Figure 1.1. Trial layout for the McGurk Effect infant habituation study.

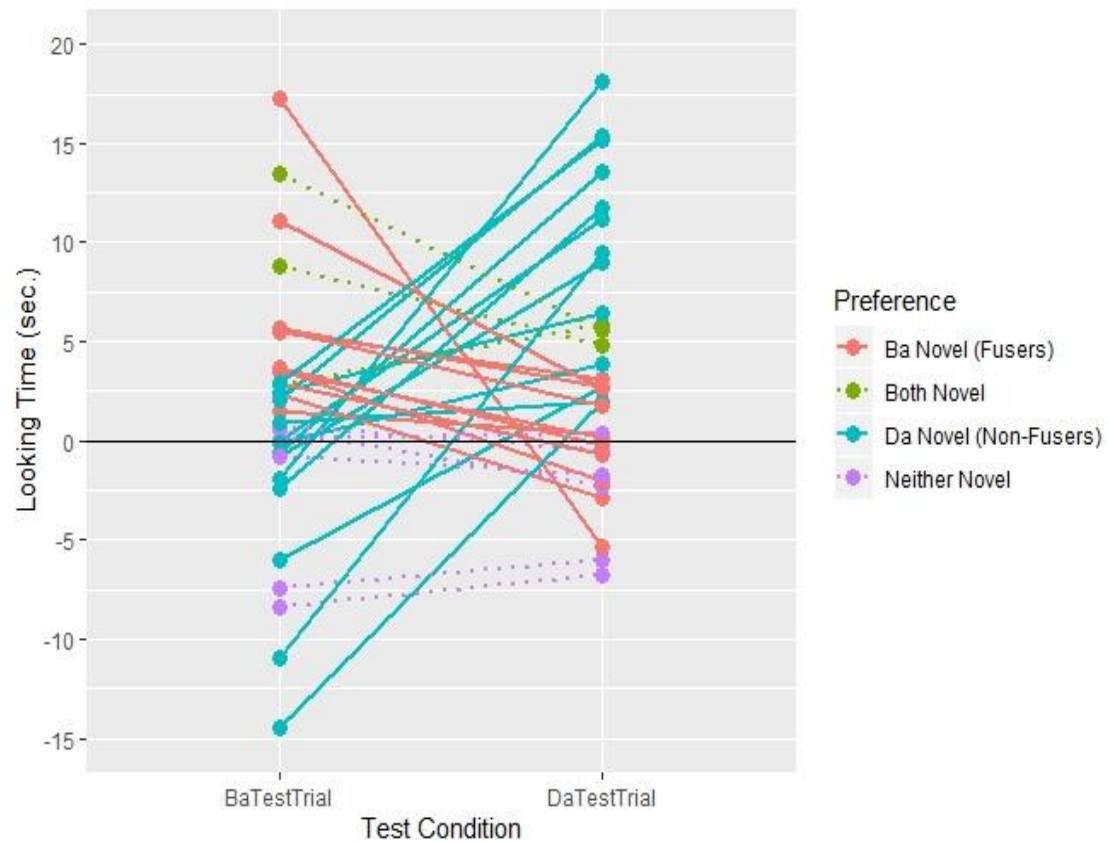


Figure 1.2. Infants' looking time difference scores for /ba/ and /da/ test trials to demonstrate individual differences in infants' McGurk Effect perception, clustered by looking behavior.

/da/ was novel as auditory perceivers (n=12). Additionally, three infants treated both /ba/ and /da/ as novel, suggesting that these infants may have perceived an alternate fusion percept (e.g., /tha/) or a combination percept (e.g., /bga/, /gaba/) and seven infants did not treat either test stimulus as novel, which may be attributed to typical, noisy infant data.

Discussion

The results from these behavioral data suggest that, as a group, the infants tested did not perceive the canonical fusion McGurk Effect. Rather, our data suggests that the current sample of infants perceived the auditory /ba/ stimulus during presentations of the canonical McGurk stimuli, evidenced by the statistically significant dishabituation to /da/ test trials. This pattern of results fails to replicate previous findings by Burnham and Dodd (1996; 2004) who found that infants perceive the fusion effect using habituation methods. Additionally, these results fail to conceptually replicate other studies of infants' perception of the McGurk Effect that have used event-related neural measures (Kushnerenko et al., 2008; Kushnerenko et al., 2008; Ujiie, Kanazawa, & Yamaguchi, 2020) or eye-tracking in an Intermodal Looking Paradigm (D'Souza et al., 2016; Mercure et al., 2019; Tomalski et al., 2013).

The discrepancies between the results presented here and the previous findings in the literature are interesting for several reasons. First, since infants were given test trials that measured their preference for both the potential fused /da/ percept and the auditory /ba/ percept, this experimental design choice afforded the ability to measure infants' individual differences in McGurk percepts. I categorized infants in four groups based on their looking times during the two test blocks: infants who were fusers, infants who were non-fusers, infants who dishabituated in both blocks, and infants who did not dishabituate in either test block. While it could be the case that, rather than individual differences, the distribution of infants' looking preferences could be due to experimental noise, pilot data from adult participants using the same McGurk Effect stimuli suggests that the distribution of results for infants is comparable to the adult pilot data (Cannistraci, Hay, & Buss, 2019). Of the 26 adult pilot participants, 46% reported perceiving the McGurk Effect (compared to 33% of infant fusers) while remaining 54% of adults were non-fusers (compared to 37% of infants). While the infants expectedly demonstrate lower rates than adults, and this difference is likely due to experimental noise in infant testing paradigms, there is a similar proportion of infants and adults who are either fusers or non-fusers. When comparing the infants to the adults, it may be the case that the noise in the infant measures manifests in the infants who did not show a preference for either test block, or the infants who dishabituated to both. However, the comparisons of the infant and adult distributions seem to suggest that individual differences interpretation of McGurk percepts is robust and present across the lifespan.

Previous work using an Intermodal Looking Paradigm to assess infants' perception of the McGurk Effect (Kushnerenko et al., 2008; Kushnerenko et al.,

2008; D'Souza et al., 2016; Mercure et al., 2019; Tomalski et al. 2013) has found that evidence across multiple studies that infants perceive the McGurk Effect. In these Intermodal Looking studies, researchers provided infants with two faces visually articulating a speech sound that were temporally synced to a single auditory syllable presentation. The researchers indexed infants' McGurk preference by their first sustained look to either face. For example, auditory /ga/ was played with the option of either congruent /ga/ or incongruent /ba/ visual information. Infants demonstrated a preference to look at the incongruent visual information. While the results from the Intermodal looking studies suggest that infants perceive the McGurk Effect, this paradigm, findings, and subsequent interpretation are different enough from the current experiment for a meaningful comparison of results between these two methods.

Future work on infants' perception of the McGurk Effect might focus on the stability of an individual's illusory perceptions of audiovisual speech. In adult participants, Mallick, Magnotti, and Beauchamp (2015) found that the test/re-test reliability of participants' perception of the McGurk Effect was stable across a year's time ($r=0.91$), however, it is an open question whether an infant's perception of the McGurk Effect (either fusion or auditory perception) stays stable across early development. The evidence suggesting stability in adults' perception of the McGurk Effect (Mallick et al, 2015) is developmentally interesting considering other adult findings that suggest auditory and visual everyday experiences are related to measured individual differences in laboratory settings. For example, Proverbio and colleagues (Proverbio, Massetti, Rizzi, & Zani, 2016) found evidence suggesting that trained musicians are less susceptible to the McGurk Effect and Brown and colleagues (Brown, Hedayati, Zanger, Mayn, Ray, Dillman-Hasso, & Strand, 2018) found that participants who are better at lip reading are more susceptible to the fusion McGurk Effect. While evidence suggests that differences in McGurk Effect perception may be modulated through everyday perceptual experiences, there is no evidence to suggest that perceiving or not perceiving the McGurk fusion is related to congruent audiovisual or auditory-only speech perception accuracy. Our results suggest that these individual differences in the perception of the McGurk Effect have their origin early in infancy and begs the questions whether these individual differences are stable and functionally relevant for auditory and visual speech perception across the lifespan.

CHAPTER II
A META-ANALYTIC INVESTIGATION OF INFANTS'
PERCEPTIONS OF THE MCGURK EFFECT

Chapter 1 demonstrates individual differences in infants' perception of the McGurk Effect in an experimental task using a within-subjects design. These findings from Chapter 1 are inconsistent with the extant literature which suggests that infants readily perceive the McGurk Effect. While our findings suggest individual differences in infants' perception of the McGurk Effect, the extant literature suggests that infants perceive the canonical McGurk Effect, as well as other illusory percepts from other incongruent audiovisual combinations. In Chapter 2 of this dissertation, I undertook a meta-analysis of infants' perception of the McGurk Effect. A meta-analytic approach is advantageous for several reasons. First, a meta-analysis allows for calculation of an average effect size and confidence interval of effect sizes that encompass both published and non-published data sets. Additionally, the meta-analysis quantitative tools can be used to examine whether publication biases or potential questionable research practices may be influencing cumulative effect size measures. Lastly, average effect sizes for subgroups of records across sources of heterogeneity in the records sampled (e.g., stimulus type or experimental paradigm) can be quantified.

In addition to computing mean effect size, effect size confidence intervals, and standard error in the meta-analysis, I was interested in quantifying three sources of heterogeneity in the infant McGurk Effect record. First, studies utilize different experimental paradigms such as habituation, intermodal looking, as well as neural measures which all have different operational definitions of the dependent variable. Studies also differed in their use of between vs within participants manipulations. Within participants designs are often more robust because multiple measures are taken from the same sample. Lastly, of the papers included in the meta-analysis, there are multiple definitions of what constitutes a McGurk Effect. Some records (e.g., Burnham & Dodd, 1996, 2004; Chapter 1) treat the McGurk Effect as the canonical fusion that occurs with auditory /ba/ and visual /ga/ stimuli, which elicits a fusion /da/ percept. Other studies (e.g., Kushnerenko et al., 2008, Kushnerenko et al., 2008; Mercure et al., 2019; Tomalski et al. 2013) use the /ba/ and /ga/ stimuli in reverse-modality composition, eliciting the McGurk Effect with an auditory /ga/ stimulus and a visual /ba/ stimulus. While this stimulus composition can indeed elicit the fused /da/ percept, it does so at a lesser rate than the canonical McGurk Effect in adult participants and is also susceptible to combination percepts (e.g., /bga/, /gaba/; see Cannistraci, Hay & Buss, 2018; Chapter 3). Other studies intended to measure the McGurk Effect in infancy used stimuli that do not elicit the canonical fusion McGurk Effect. Desjardins and Werker (2004) as well as Rosenblum, Schmuckler, and Johnson (1997) use stimuli that elicit a perceptual effect where the visual speech information overrides the auditory input. In these cases, an auditory /ba/ is dubbed with a visual /va/, and infants and adults alike perceive the visual /va/.

Differences in stimuli choice, as well as other experimental considerations, may influence the effect size and variability from infant McGurk Effect studies.

Meta-analyses are equipped to collate these independent empirical findings into a quantitative summary of how robustly infants perceive the McGurk Effect. Heterogeneity exists across the scientific record because individual studies are approached with specific research goals in mind. Thus, experiments often must balance the utility and limitations of their experimental designs. In the current meta-analysis, I aimed to calculate a meta-analytic mean effect size for infants' perception of the McGurk Effect. Additionally, I quantified different sources of heterogeneity in the scientific record, and how these experimental design choices may influence the field's research practices and resulting publication trends.

Methods

The published studies included in the current meta-analysis were first identified by an expert list, compiled by the author of this dissertation. This expert list yielded four published empirical papers, from which I included 8 research records. On 9/7/2020, I executed a Google Scholar search using the search terms "McGurk Effect" & "Infant" OR "Child," and this search yielded 1,420 results. I examined each of the 1,420 search items and excluded papers that did not have relevant data or duplicates that were previously identified as part of our expert list. After this assessment, the search yielded 6 new published manuscripts, with one included record from each paper. I cross referenced this search using the same search terms on PubMed, which yielded 551 hits, though none of these hits were unique. I set up a Google Scholar alert with the same search terms as above, and this alert gathered one additional paper (Ujiie, Kanazawa, & Yamaguchi, 2020). However, this paper was excluded from our analyses for two reasons, the experimenters played background noise to more readily induce the canonical McGurk Effect, and the researchers also used the audiovisual speech to study infants' perception of the same and other-race faces. On 4/19/2021, I solicited two developmental psychology listservs for any published or unpublished McGurk Effect data. This call yielded one additional unpublished manuscript (Riva, 2021) that contained two records that were included in the analysis. Lastly, the empirical data presented in Chapter 1 were also included. In sum, our data set for the meta-analysis included data from 10 manuscripts and these manuscripts yielded 15 unique research records (i.e., experimental conditions included in the analysis).

I conducted the meta-analysis using R statistical software (R Core Team, 2020; meta and dmetr packages) as well as a meta-analysis guide from Harrer and colleagues (Harrer, Cuijpers, Furukawa, & Ebert, 2021). I used a random-effects-model for this meta-analysis because the included data come from a heterogeneous population of infants across different ages, cultures, and native languages. The random-effects-model approach assumes that, in addition to sampling error, the different populations from which the empirical data are collected are an additional source of variability. The random-effects model is also advantageous because it is more accurate when a smaller number of studies are included in the meta-analyses (Schwarzer, Carpenter, & Rücker, 2015). These

considerations for random-effects models are especially relevant for developmental research. I calculated the Cohen's d and Hedge's g effect size for each of the 14 records included. I used Hedge's g for the calculations in the meta-analysis since it is more accurate for studies with small sample sizes (Hedges, 1981; Morris & DeShon, 2002).

Results

I calculated that the variance-weighted meta-analytic effect using Hedge's g values for infants' perception of the McGurk Effect ($g=0.55$, $CI[0.40,0.69]$). This meta-analytic effect was different from zero, $t(15)=7.89$, $p<0.001$, suggesting a moderate true effect of infants' ability to perceive the McGurk Effect (see *Figure 2.1*). *Figure 2.2* shows a funnel plot of each included record's effect size and standard error, and I tested for the potential of publication bias using an Egger's funnel plot asymmetry test. The Egger's test for funnel plot asymmetry is marginally significant ($t(14)=1.963$, $p=0.07$) suggesting that there is trending evidence towards evidence of publication bias in the surveyed infants McGurk Effect studies. Lastly, I ran P-Curve analyses (*Figure 2.3*) on our data's right-skewness to assess the presence of potential data manipulation and a flatness test of statistical power. For the right-skewness test, evidence of right-skewness suggests that records with statistically significant p -values are distributed in an expected manner, and that there is no evidence of data manipulation or p -hacking. Evidence of left-skewed data suggests the presence of data manipulation, since a greater proportion of significant results would be artificially close to the accepted $p=0.05$ threshold. The data included in our meta-analysis are significantly right-skewed ($p=0.048$) suggesting no evidence of data manipulation or p -hacking. The P-Curve flatness test was also non-significant ($p=0.995$), suggesting that the cumulative effect size calculation has sufficient statistical power and that infants' perception of the McGurk Effect is indeed a true effect.

To probe any meaningful differences across the experimental designs of the studies included in the meta-analysis, I conducted three, follow-up subgroup analyses. Subgroup analyses pool effect sizes for specific designations of records included in the meta-analysis and lends these pooled effects to statistical comparisons. I conducted a stimulus type subgroup analysis, breaking stimulus type into three groups: McGurk Fusion (Aud/ba/ Vis/ga/), Audiovisual Mismatch (Aud/ga/ Vis/ba/ or Aud/ka/ Vis/pa/), and Visual Dominance (Aud/ba/ Vis/va/). The pooled effect size for Fusion stimuli was the lowest ($g=0.40$), followed by Visual Dominance ($g=0.51$) and AV Mismatch ($g=0.66$). However, these differences were not statistically significant ($Q=4.16$, $p=0.125$; see *Figure 2.4*). Next, I conducted a method subgroup analysis, breaking stimulus type into three groups: Habituation, Intermodal Looking, and EEG (using audiovisual mismatch negativity responses measured from left fronto-temporal electrodes). The pooled effect size for Habituation was the lowest ($g=0.47$), followed by EEG ($g=0.61$) and Intermodal Looking ($g=0.63$). These differences were also non-significant

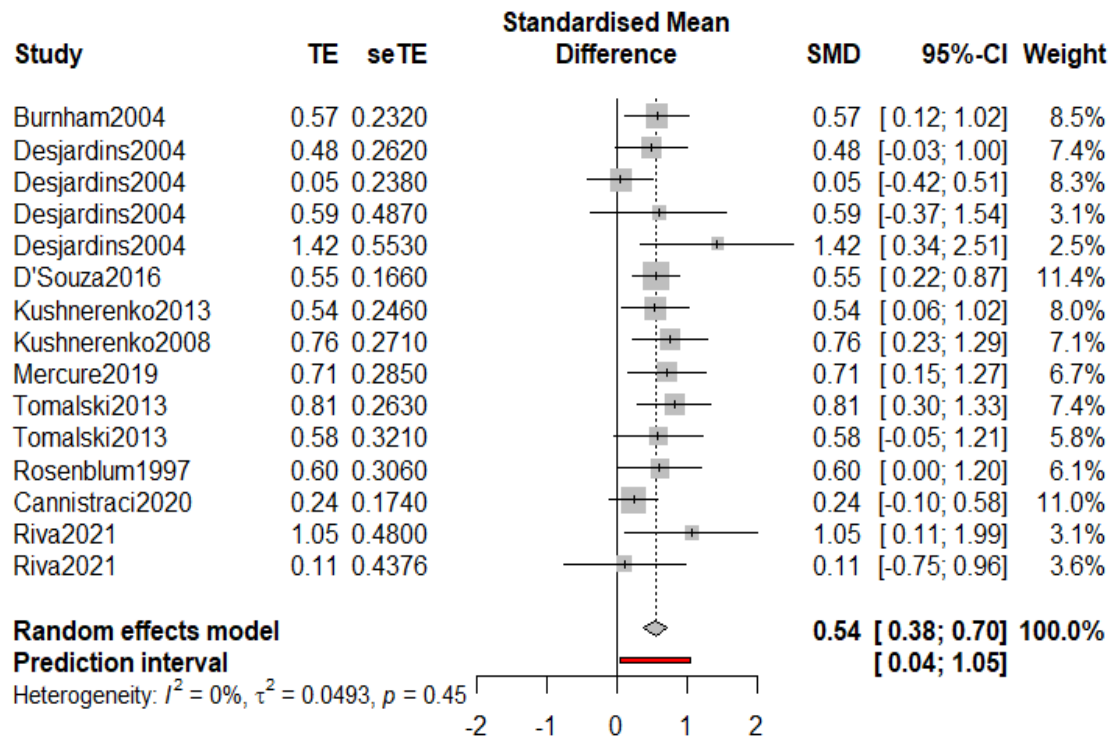
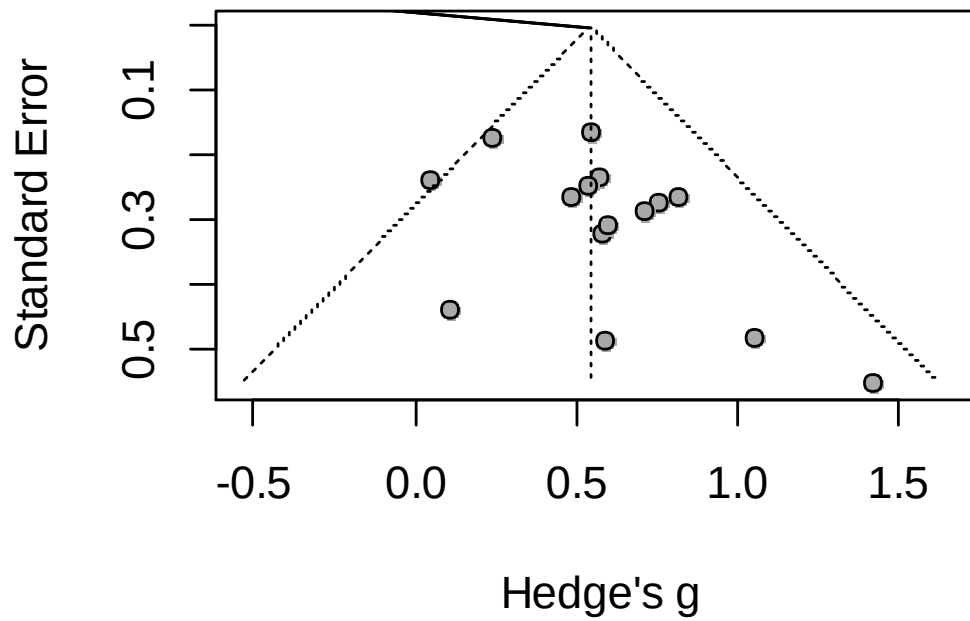


Figure 2.1. Forest plot depicting the average meta-analytic effect of the random effect model (SMD), as well as the effect size (TE), standard error (seTE), confidence interval (95%-CI), and model weighting (Weight) of each included research record.



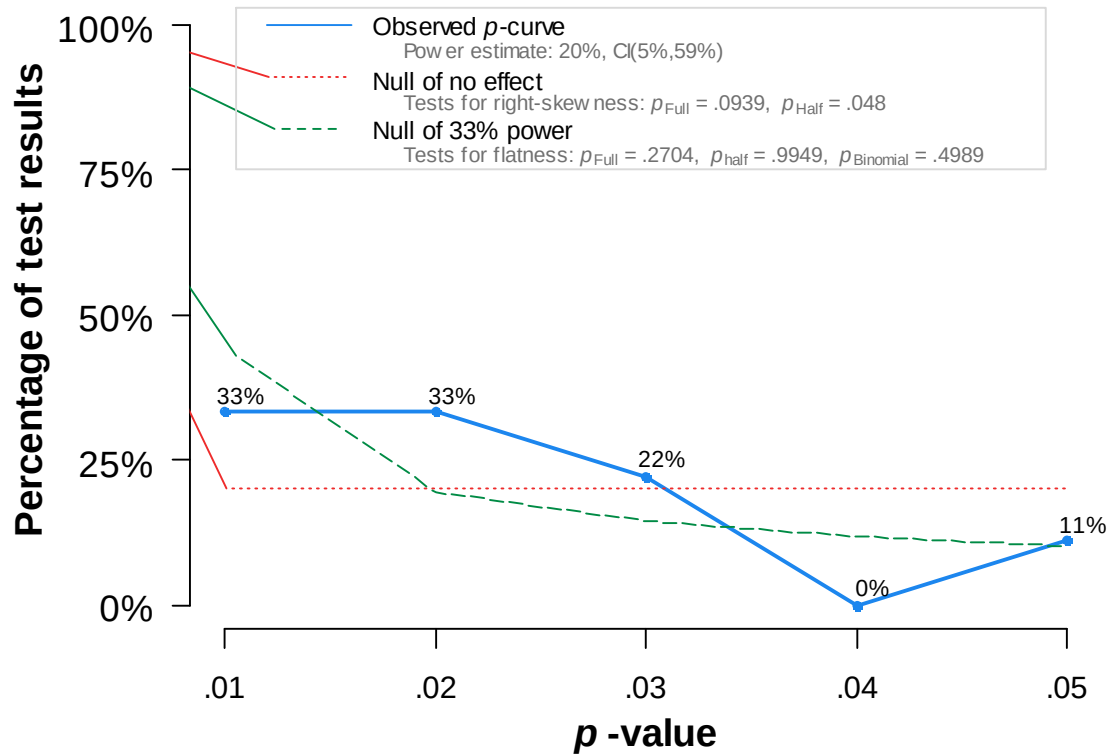


Figure 2.3. P-Curve plot including the statistically significant records from in the meta-analysis. These statistically significant records pass right-skewness and flatness tests, suggesting that the significant results are distributed in a pattern that does not suggest evidence of data manipulation and that there is sufficient power to detect these significant results.

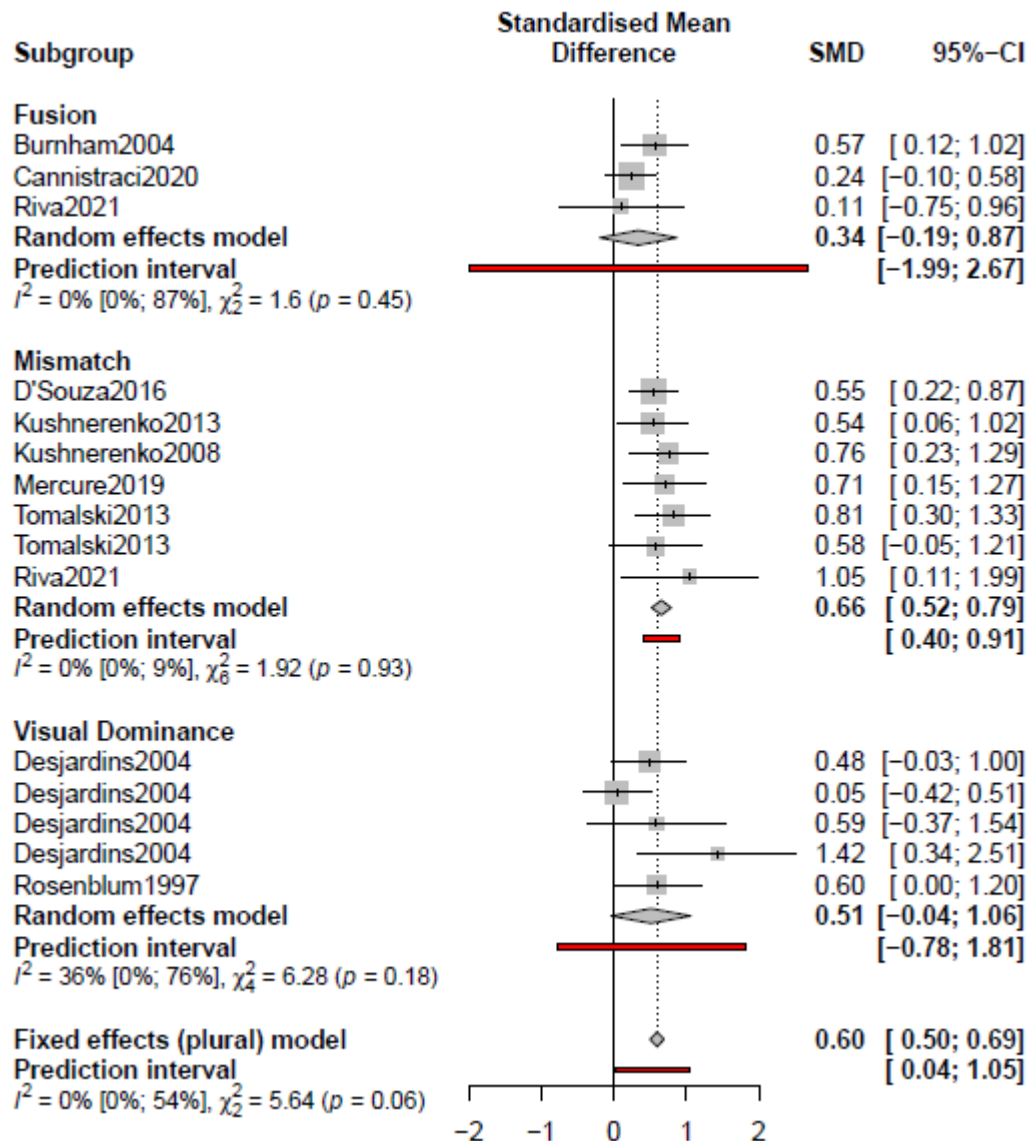


Figure 2.4. A forest plot for stimulus combination subgroup analysis with fusion stimuli at the top, audiovisual mismatch stimuli in the middle, and visually-dominant stimuli at the bottom. SMD refers to the mean effect size for each of the subgroups and the prediction interval spans the range where effect sizes using these stimuli combination would be expected to be normally distributed.

($Q=1.48$, $p=0.477$; see *Figure 2.5*). Finally, I conducted a design subgroup analysis (Between Subjects vs Within Subjects). Similarly, there were no statistically significant differences in Between Subjects ($g=0.53$) vs Within Subjects ($g=0.56$) experimental designs ($Q=0.04$, $p=0.845$; see *Figure 2.6*). Taken together, while there are differences in the stimuli, paradigm, and experimental design across the included studies, there are no statistically significant differences that arise due to the measured between-studies variability.

Discussion

The meta-analysis presented here focused on quantifying infants' perception of the McGurk Effect across the scientific record. The results of this analysis suggest a moderate, statistically significant meta-analytic effect of infants' perception of the McGurk Effect. Interestingly, there is trending evidence to suggest there may be some publication bias in the infant McGurk Effect literature. Future additions to this meta-analysis from the unpublished records will continue to quantify whether significant results providing evidence of infants' perception of the McGurk Effect are disproportionately present in the scientific record. In the current analysis, there were no significant effects of sub-groupings within the literature, such as experimental design (between vs within subjects) or the testing paradigm used (Habituation, Intermodal Looking, EEG). While there were also no significant differences between different stimulus combinations used in the literature, the differences in effect sizes are qualitatively interesting in the scope of the behavioral data from Chapter 1. Studies that tested the McGurk effect using canonical auditory /ba/ and visual /ga/ stimuli had the smallest cumulative effect size ($g=0.34$), while studies using visually dominant stimuli ($g=0.51$) or audiovisual mismatch ($g=0.66$) stimuli showed larger effect sizes. While there were no significant differences in the subgroup analysis for stimulus combination, the stimuli used in Chapter 1 are consistent with the lowest measured effect size in the meta-analysis.

The differences in defining the McGurk Effect may be especially relevant for future research, especially when interpreting how the McGurk Effect is related to language outcomes. In the introduction section of this dissertation, I argued that infants' perception of the McGurk Effect is evidence that they can integrate auditory and visual speech information. This interpretation is sensible for stimulus combinations that involve the canonical fusion effect, since the percept is a syllable that is not present in the auditory nor visual speech information. For example, it might be the case that the stimuli used in intermodal looking studies of the McGurk Effect in infants (e.g., D'Souza et al., 2016; Kushnerenko et al., 2008, Kushnerenko et al., 2013; Mercure et al., 2019; Tomalski et al. 2013), might induce a combination percept (e.g., /bga/ or /gba/). These studies interpret infants' first look in a trial towards the incongruent face as McGurk Effect. Infants with this pattern of responding demonstrate a significant audiovisual mismatch negativity response in EEG studies (e.g., Kushnerenko et al., 2013). It is possible, however, that these behavioral and neural responses might be

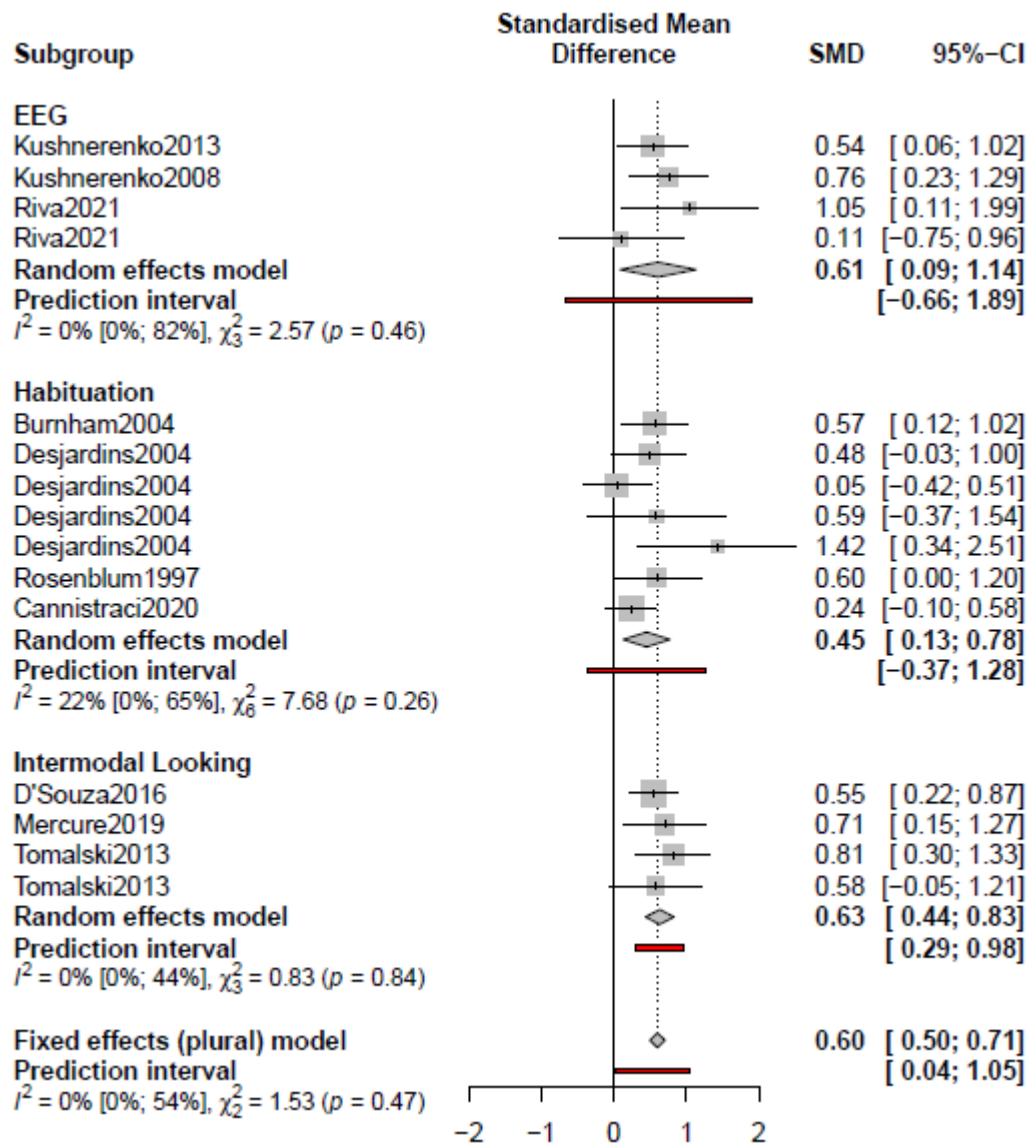


Figure 2.5. Forest plot for experimental design subgroup analysis, including EEG methods at the top, Habituation methods in the middle, and Intermodal looking methods at the bottom. SMD refers to the mean effect size for each subgroup and the prediction interval is the expected range for normally distributed effect sizes for future work using the three experimental methods.

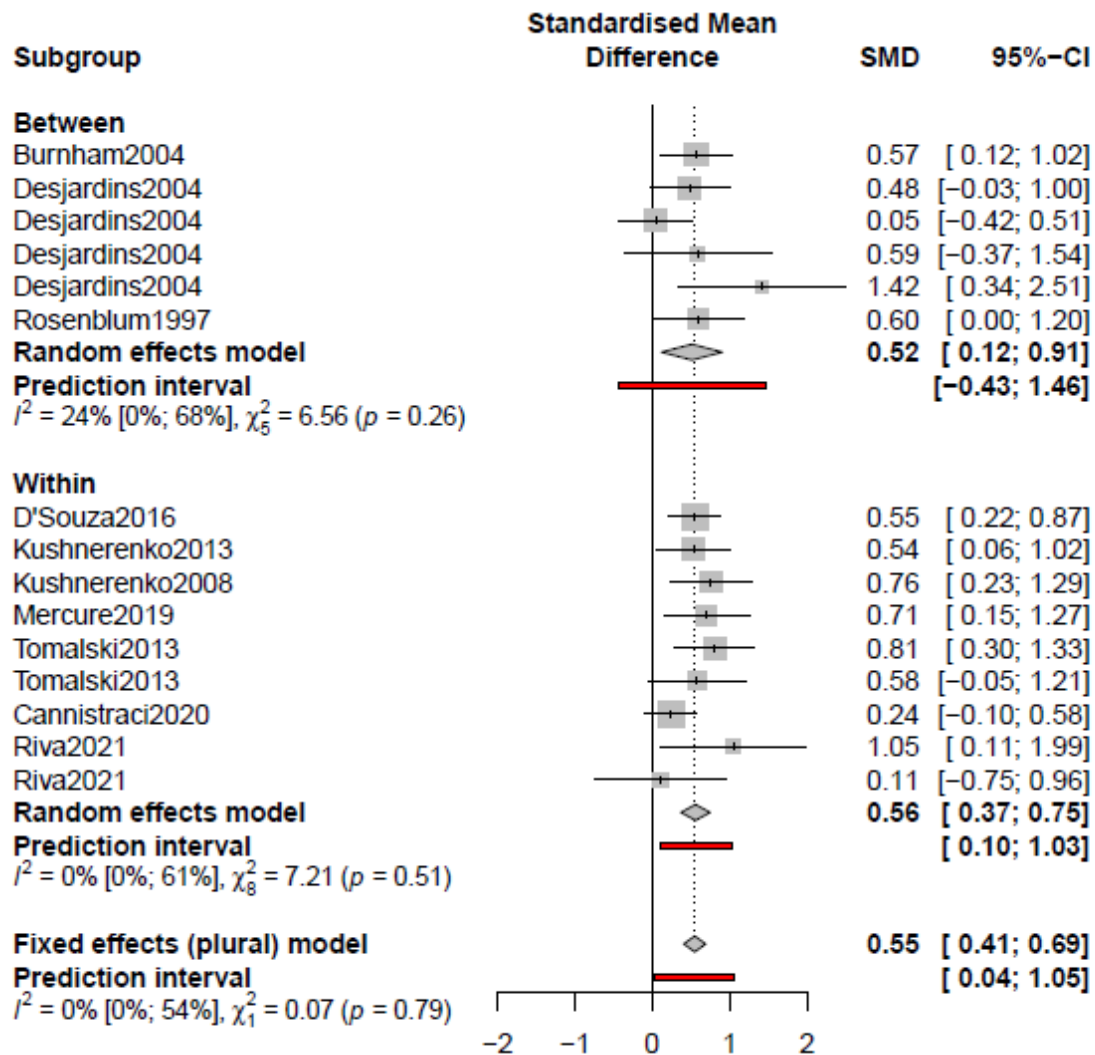


Figure 2.6. Forest plot for between (top) vs within-subjects (bottom) subgroup analysis. SMD is the mean effect size for each subgroup and the prediction interval spans the expected range wherein effect sizes may be normally distributed.

indicative of a combination percept rather than a fusion McGurk Effect. In Riva (2021), the canonical McGurk Effect condition was not statistically different from audiovisually congruent controls, though infants' audiovisual mismatch responses were significantly different from chance for the auditory /ga/ and visual /ba/ stimuli that may induce combination percepts. A similar pattern of results was found by Kushnerenko and colleagues (Kushnerenko et al., 2013), where only the auditory /ga/ and visual /ba/ stimulus elicited the audiovisual mismatch negativity response.

In future work, researchers should be explicit in describing the stimuli and the methods used to quantify measuring the McGurk Effect in infancy. Of the work considered in the meta-analysis, only one study (Burnham & Dodd, 2004) provided evidence that infants perceive the McGurk fusion. In a developmental context, an infant's perception of the McGurk Effect, as broadly defined in the meta-analysis inclusion criteria, could be referring to a fusion effect, a visually dominant effect, or a combination percept. Each of these audiovisual illusory effects may be differentially related to language outcomes. I hypothesized that the canonical McGurk Fusion effect would be correlated with later vocabulary outcomes, but it could be the case that visually-dominant illusions or auditory preferences might be better predictors. To further situate these concerns, I aimed to study differences in perceptual processing for a range of audiovisual illusions in Chapter 3 and explore the potential longitudinal correspondences between audiovisual speech perception at 4.5 months and later language outcomes in Chapter 4. In sum, the prudent conclusion from our meta-analysis results is that infants can perceive illusions for audiovisual speech stimuli—however, researchers should be cautious and transparent about the specific stimuli used to induce the canonical McGurk Effect, or other audiovisual speech illusions.

CHAPTER III
A DYNAMIC NEURAL FIELD EXPLORATION OF INDIVIDUAL
DIFFERENCES IN THE MCGURK EFFECT

The empirical results from Chapter 1 suggest that, as a group, infants in our experiment perceive the auditory /ba/ stimulus when viewing our McGurk stimuli. However, there were substantial individual differences in our sample of 4.5-month-olds. Further, meta-analytic evidence from Chapter 2 suggests that, across the scientific record, infants show a moderate evidence of the McGurk Effect. One potential tool that can be used to further assess both trends in the literature and individual differences in audiovisual speech perception is *Dynamic Neural Field (DNF)* computational modeling. DNF models are a class of dynamical systems models (Schöner & Spencer, 2016) that have been previously used to elucidate how brain-behavior relationships function to support real-time behavior in adults, children, and infants (Buss, Wifall, Hazeltine, & Spencer, 2014; Buss & Spencer, 2014; Perone & Spencer, 2013). The computations used in DNF models are grounded in the measured neurophysiology of neural field dynamics (e.g., Amari, 1977; Georgopoulos, Schwartz, & Kettner, 1986; see the Appendix for the specific computations used in Chapter 3).

DNF models afford an embodied framework to formally model how audiovisual speech perception might transverse behavioral and neural levels of analysis. Previous computational models of speech perception have aimed to describe audiovisual speech perception in the terms of Connectionist (Vallabha, McClelland, Pons, Werker, & Amano, 2007), Causal Inference (Magnotti & Beauchamp, 2017), and Mix of Gaussian (Getz, Nordeen, Vrabic, & Toscano, 2017) models. Each of these modeling frameworks (Connectionist, Causal Inference, and Mix of Gaussian) incorporate competitive, winner-take-all computational processes to guide the formation of phoneme representations and the perceptual process of auditory and/or audiovisual speech. The DNF model presented in this chapter also assumes that neural representations of auditory and visual speech information are acquired over perceptual experience (instantiated by Hebbian learning traces). Critically, the DNF modeling framework utilizes competitive neural processes of local excitation and lateral inhibition, rather than Bayesian computation, and these neural dynamics are coordinated in real-time to support emergent perceptual phenomena, in this case, speech perception.

Using the DNF model presented here in Chapter 3, I identify and manipulate crucial, neurally-plausible parameters to demonstrate how individual differences in audiovisual speech perception may emerge and lead some people to perceive the McGurk Effect, while leading others to perceive the auditory /ba/ stimulus. I first walk through a full simulation of illusory audiovisual speech illusions and an auditory-only fusion found in the extant literature, as well as auditory-only and visual-only controls in an adult model simulation to demonstrate the utility of this approach. I next embed an infant version of the model in a habituation paradigm that replicates the experimental design of our behavioral data in Chapter 1. This framework allows us to investigate how

individual differences in neural and perceptual processes might give rise to the measured individual differences found in Chapter 1.

Methods

Model Architecture

I constructed a feature integration DNF model (e.g., Schöner & Spencer, 2016) using the COSIVINA toolbox in Matlab (Schneegans, 2015). The COSIVINA package contains a coding environment outfitted with the relevant mathematical functions and interaction kernels commonly used in DNF models (Schöner & Spencer, 2016; see Appendix for specific equations used here). The model is composed of five neural fields that bind auditory and visual speech information with neural fields that support perceptual and working memory processes. First, audiovisual speech stimuli must be bound to neural representations of the speech sounds and orofacial movements used to articulate speech. To accomplish this, two separate two-dimensional excitatory fields bind incoming auditory and visual sensory input with Hebbian learning traces of previously experienced syllable combinations. The two-dimensional visual and auditory fields capture visual and auditory speech in neural fields organized by perceptual distinctiveness. That is, auditory traces that share similar features are closer in space to one another (e.g., voiced consonants are clustered), and the visual traces are clustered based on the similarity of the visual articulatory information (*Figure 3.1*). For example, the traces for visual /ba/ and /va/ are closer together because they are both articulated using the lips, while /ba/ and /ga/ are more distant since /ga/ is a velar sound. The model includes Hebbian learning traces for 5 syllable combinations across all the neural fields (/ba/, /va/, /tha/, /da/, & /ga/) that are relevant for perceiving the McGurk Effect, as well as the other audiovisual illusory effects outlined in the meta-analysis in Chapter 2.

The two-dimensional auditory and visual fields are recurrently linked to an excitatory perceptual field (the “phoneme field”) that binds sensory information from the auditory and visual inputs to give rise to the model’s perceptual output. Secondly, a working memory field builds and maintains a working memory peak of the model’s perceptual experience. A lateral inhibition field is connected to both the phoneme and working memory fields, which balances local excitation within each field and lateral inhibition to give rise to stable patterns of activation across these three fields. The three-layer perceptual, inhibitory, and working memory architecture is quite common in DNF models (Schöner & Spencer, 2016). The only difference between the adult and infant models are that in the infant model, output from the perceptual and working memory layers are linked to an autonomous looking system (e.g., Perone & Spencer, 2013), composed of attention and working memory nodes that simulate the model’s looking behavior during habituation (*Figure 3.2*). The attention node receives activation from the perceptual field, such that when the perceptual field is forming a peak, activation from this neural process excites the attention node and this simulates the model

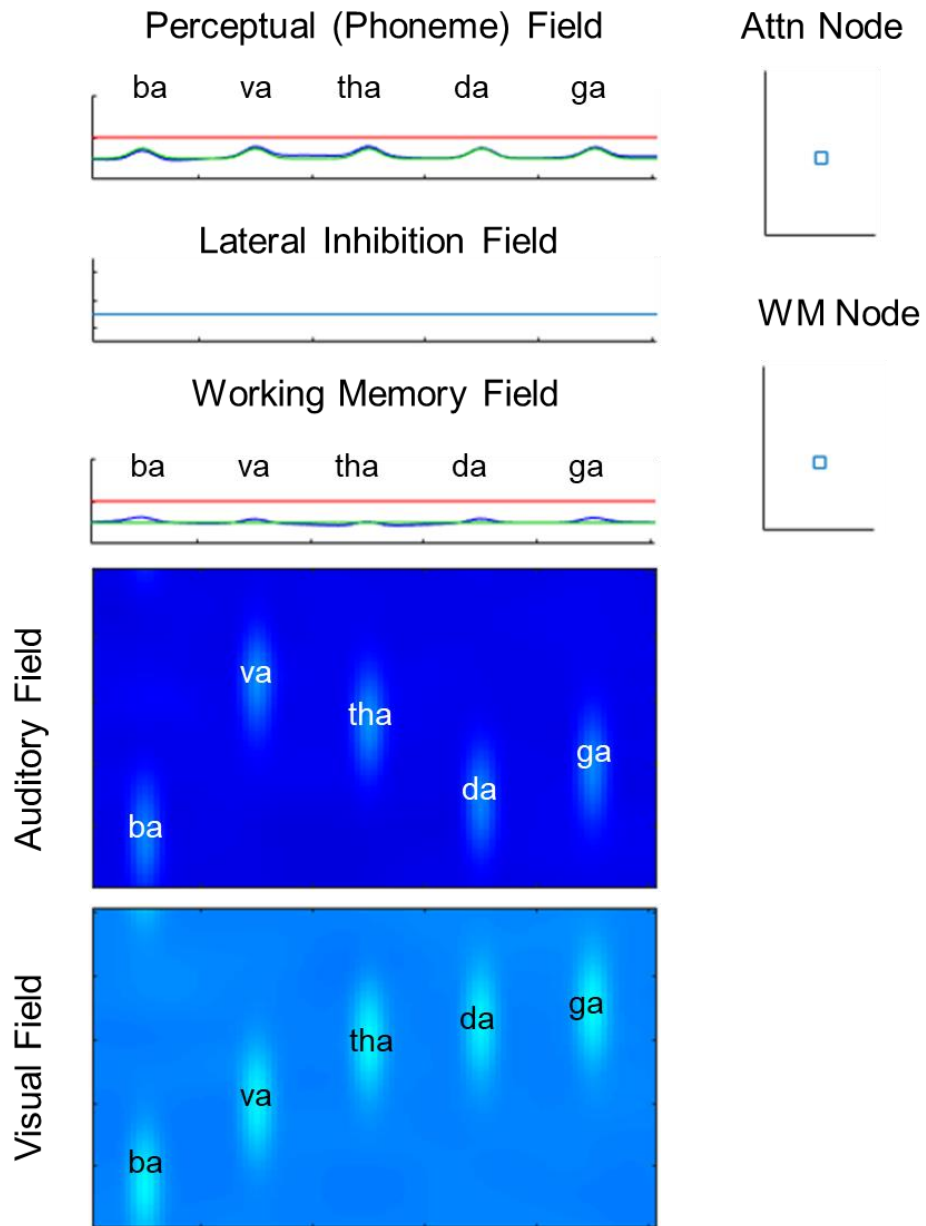


Figure 3.1. The DNF model architecture of the infant McGurk Effect model at resting state. Hebbian learning pre-shapes for phonemes across each field are labeled.

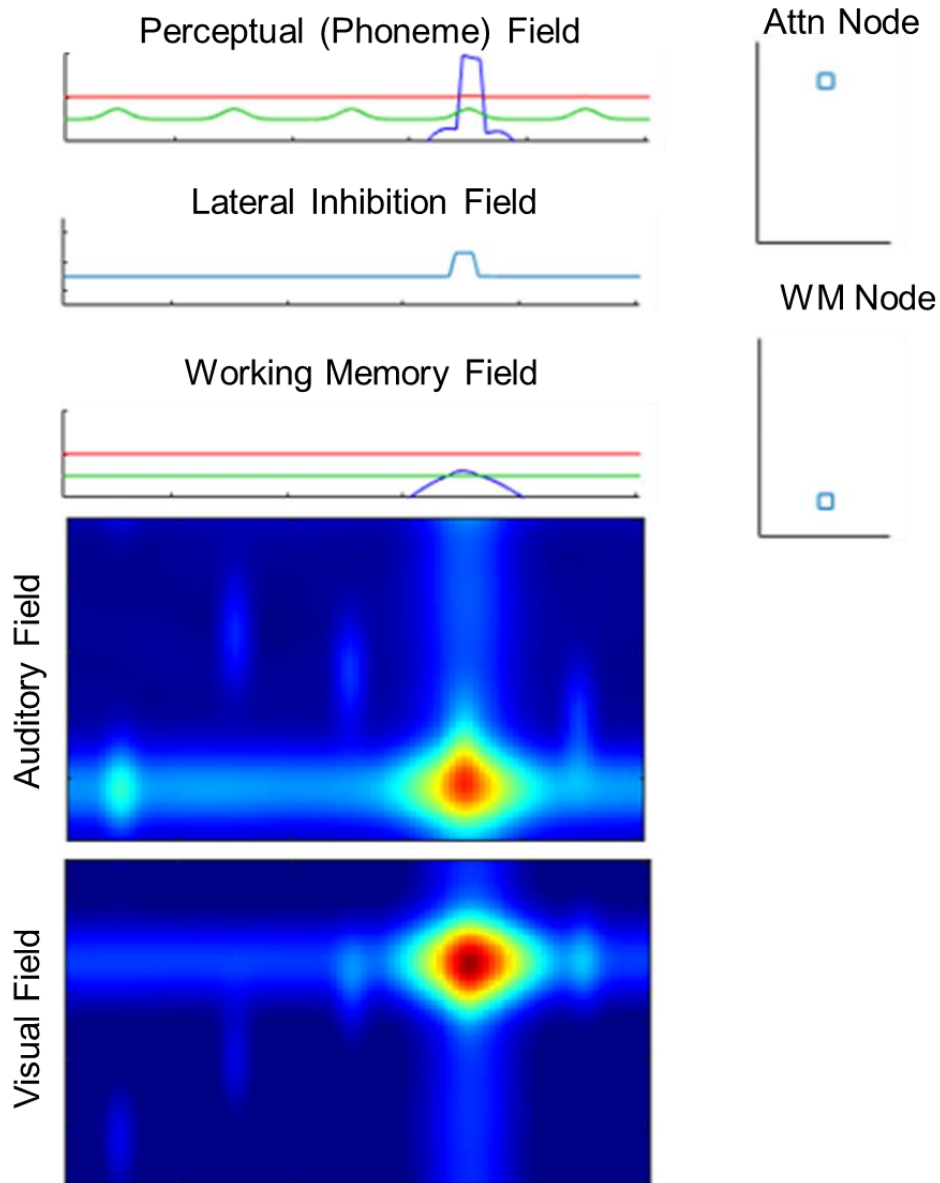


Figure 3.2. The DNF model architecture of the infant McGurk Effect model binding the fusion effect in the Perceptual Field, positive activation in the Attention Node, and a working memory peak forming below threshold in the Working Memory Field.

attending to the stimuli it is experiencing. This pattern of responding is interpreted as an infant sustaining attention to looking at stimulus during a real-life habituation experimental task. The working memory node receives excitatory activity from the working memory layer, and when a working memory peak is formed, the excitation that is recurrently coupled with the working memory node increases its polarity and simultaneously begins to inhibit the excitatory activity in the attention node. This pattern of responding is meant to simulate an infant becoming habituated to the stimuli in the task, and the model then breaks its selective attention and looks away from the experimental stimuli. These looking dynamics are central to infant habituation measures and the DNF modeling framework allows researchers to simulate how these attention and working memory dynamics might support infants' looking behavior in real-time in habituation tasks.

Model Simulations

Both the adult and infant model simulations were run in Matlab. The adult model was couched to simulate literature-based rates, since I was interested in uncovering how often each of the illusory effects were found in the literature. Thus, I ran 100 trials of each stimulus combination for the following conditions to compute a mean rate of each effect: (1) audiovisual congruence (auditory /ba/, visual /ba/), (2) visual-dominant percepts (e.g., Rosenblum & Saldaña, 1996: auditory /ba/, visual /va/), (3) audiovisual fusion percepts (e.g., McGurk & MacDonald, 1976: auditory /ba/, visual /ga/), (4) auditory-dominant percepts (e.g., MacDonald & McGurk, 1978; Magnotti & Beauchamp, 2017: auditory /ga/, visual /ba/), (5) dichotic fusion (e.g., Cutting, 1976: auditory /ba/, auditory /ga/), (6) auditory-only (/ba/), and (7) visual-only (/ba/). In addition to the canonical McGurk Effect and the audiovisual, auditory-only, and visual-only controls to ensure our model accurately perceives speech across modality combinations, I was interested in modeling the rate at which adult participants perceive auditory-dominant effects in the reverse coded McGurk Effect, since those stimuli have been used in intermodal looking paradigms with infants (see meta-analysis in Chapter 2). Additionally, there is evidence that an auditory-only fusion effect (Cutting, 1976) can occur a dichotic listening tasks from auditory /ba/ and /ga/ being presented in either ear of the participant, so I tested whether our model architecture could also simulate this effect from simultaneous stimulus presentation of auditory /ba/ and /ga/.

In the infant task, I aimed to replicate the infants' experience from the behavioral experiment conducted in Chapter 1. Infants were habituated to the canonical McGurk Effect (i.e., auditory /ba/, visual /ga/) and the model's "looking behavior" was interpreted from the activity of the attention and working memory nodes, similar to Perone and Spencer (2013). I used the same habituation criterion as the behavioral experiment in Chapter 1; criterion was calculated by averaging the model's attention node activity (as an analogue of infants' looking times) for the first three habituation trials. I compared a 3-trial averaged sliding

window of each model's "looking time" to the looking criterion. In effect, the looking behavior in the models, as well as time and trials to habituate were "infant controlled" as in Chapter 1, with the looking dynamics of the model controlling stimulus presentations.

Based on the behavioral results from Chapter 1, I ran 40 total simulations: 20 simulations were run using a model tuning that robustly preferred the fusion percept from McGurk stimuli, and the remaining 20 simulations were run with a model tuning that preferred the auditory /ba/ percept. Both the "fuser" and "auditory preference" model were hand-tuned variants from the initial parameters of the adult model. The "fuser" model tuning had greater self-excitation and lateral inhibition parameter values in the auditory and visual 2-dimensional fields, as well as slightly larger global inhibition in the visual field, to help temper the increase in local excitation. I tuned these local field dynamics to allow for greater excitatory neural processes towards the /da/ fusion Hebbian pre-shapes, which facilitated fusion percepts. To test the hypothesis that individuals with weaker visual-to-phoneme field coupling would be less susceptible to the McGurk Fusion Effect, the "auditory preference" model had a slightly weaker projection from the visual 2-dimensional field to the perceptual (phoneme) field. In future work using this modeling architecture, I plan to instantiate the same test blocks and refamiliarization trials as in Chapter 1 after habituation to and record each model's "looking time" for these test trials.

Results

The adult McGurk Effect model was able to capture audiovisual speech effects reported in the literature as well as congruent controls (*Figure 3.3*). For control conditions, the model (M) and literature (L) values were quite close: audiovisual congruence (M=99%, L=100%), auditory-only (M=90%, L=100%), and visual-only (M=100%, L=100%). For audiovisual illusory effects, the model outperformed some of the rates in the literature but was within at least a 20% interval from the rates culled from the extant literature and the rates were as follows: visual-dominance (M=81%, L=60.5%), McGurk fusion (M=93%, L=81%), and auditory-dominance (M=87%, L=71%). Additionally, extant literature suggests that some of the audiovisual illusions produce stable alternative percepts, and the simulation also captured these percepts (McGurk & MacDonald, 1976; Getz & Toscano, 2021). For the auditory-dominance condition, a combination percept (e.g., /baga/) formed when stable peaks were maintained at both /ba/ and /ga/ locations in the perceptual, phoneme field 8% of the time. Lastly, dichotic fusion rates were also close (M=78%, L=68%) to the literature (Cutting, 1976), with the prevailing alternative percept also being a /baga/ combination effect (14%).

For the infant McGurk model tuned to perceive the fusion percept, 4 simulations did not have looking dynamics that met the habituation criteria. The 16 remaining simulations all showed evidence of perceiving the McGurk fusion percept. Of the 20 simulations tuned to perceive the auditory /ba/ percept, 1

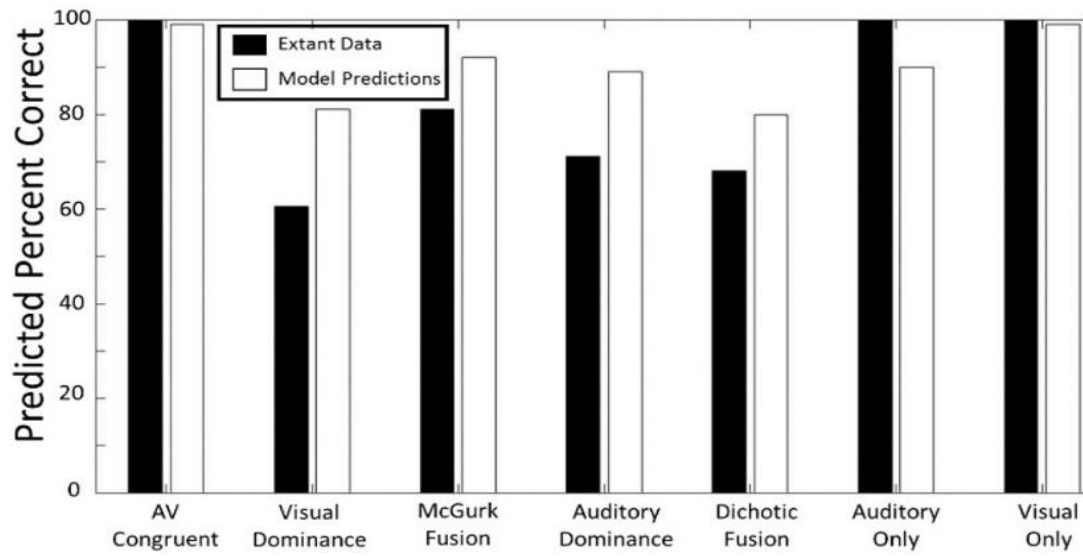


Figure 3.3. Adult McGurk Effect model simulation results compared to observed rates in the extant literature.

simulation did not meet habituation criteria, 1 simulation perceived the fusion effect, and 1 simulation perceived a visual /ga/ percept—however, the remaining 17 simulations indeed perceived the auditory /ba/ percept. When averaging across the 40 simulation runs, the models took an average of 14 trials to reach habituation criteria, while 5 simulations in total did not meet habituation criteria (87.5% inclusion rate). In the behavioral data collected in Chapter 1, infants averaged 9 trials to habituate and 12 infants were excluded for fussiness or failure to habituate (73.3% inclusion rate). Of course, model simulations do not face the same attrition and fussiness challenges as 4.5-month-old infants, and while the number of trials to habituate and the inclusion rates are larger for the model simulations, they are reasonable values for infant research.

Discussion

The adult and infant McGurk Effect simulations presented here demonstrate how an architecture of neural fields can account for audiovisual speech perception across different tasks and stimulus combinations. The adult model simulates a wide range of perceptual effects across both auditory and visual domains at a similar rate found in the extant literature. Current computational models of the McGurk Effect focus solely on the canonical fusion effect with auditory-only or congruent audiovisual speech as controls. Simulating the canonical McGurk Effect is a good start, our DNF model suggests that the future computational work should be able to simulate a wide range of effects—Massaro (2017) makes a similar point in his recent commentary. While the previous literature uses the term “McGurk Effect” quite liberally in reference to different combinations of incongruent audiovisual speech, it might be helpful for future work in the field to specify what specific outcome is most common, for example, visually-dominant effects or auditory-fusion effects.

The infant McGurk model was created to assess infants’ individual differences in the perception of the McGurk Effect. I identified relevant tuning parameters across neural field connections, such as weaker projections from sensory fields (i.e., visual field) to perceptual layers, as well as altering the local field dynamics to create more/less discrete binding in sensory processing. Infants’ perception of audiovisual speech information may influence later language outcomes, so utilizing DNF models to identify drivers of developmental change and individual differences could benefit our understanding of language development (see Buss & Spencer, 2018 for a similar approach for the development of Executive Function). Indeed, Buss and colleagues (Buss & Spencer, 2018; Buss, Magnotta, Penny, Schöner, Huppert, & Spencer, 2021) have used DNF models to predict hemodynamic responses of neural fields, in addition to predicting behavioral outcomes on experimental tasks. These DNF models can therefore act to assess brain-behavioral relationships across development and give perspective to how neural and behavioral dynamics change across development. Future work on infants’ audiovisual speech perception and computational modeling should similarly attempt to bridge how

relevant neural dynamics might support language learning. Samuelson and colleagues (Bhat, Spencer, & Samuelson, 2018; Samuelson, 2021) have taken DNF approaches to word learning and cross-situational learning, and thus the future of DNF speech perception models might aim to incorporate speech perception dynamics with larger questions of how infants' perceptual and cognitive development scaffolds their language learning.

CHAPTER IV
LONGITUDINAL CORRESPONDANCES BETWEEN INFANTS'
AUDIOVISUAL SPEECH PERCEPTION AND NATIVE LANGUAGE
PERCEPTUAL NARROWING

Chapter 4 aims to study longitudinal relationships between early audiovisual integration and how this multimodal processing ability may be functionally related to later language outcomes. Despite evidence in the extant literature and the findings from the first two chapters suggesting audiovisual speech information may be especially ecologically relevant for infants, most studies on the development of speech perception skills have been limited to infants' processing abilities of the auditory speech signal. At birth, human infants show evidence that they can discriminate many phonemes (Eimas, Siqueland, Jusczyk, & Vigorito, 1971) and within the first year of life, infants begin categorizing the speech sounds of the native language(s) (Kuhl, Conboy, Coffey-Corina, Padden, Rivera-Gaxiola, & Nelson, 2008). This native language attunement process is twofold, encompassing the infant's ability to home in on the relevant contrasts in their native language(s), while also becoming less sensitive to non-native contrasts (Werker & Tees, 1984; Kuhl, 1991).

Interestingly, previous research suggests that some phoneme contrasts are more difficult to discriminate than others. For example, 3-month-old infants show evidence for better discrimination of phoneme contrasts that vary in place-of-articulation (e.g., /b/ vs /d/) compared to phoneme contrasts where voicing changes, but place-of-articulation remains constant (e.g., /g/ vs /k/: Eimas, 1974). Kuhl and colleagues (Kuhl, Stevens, Hayashi, Deguchi, Kiritani, & Iverson, 2006) demonstrated that accuracy on some place-of-articulation distinctions is influenced by native language phoneme distributional probabilities. Infants learning English as their native language become increasingly sensitive to native language /r/ vs /l/ contrasts, while Japanese learning infants lose sensitivity to the same contrast. Kuhl and colleagues hypothesized that this is the case because English /r/ and /l/ are not contrastive in Japanese and are instead collapsed into a single phonological category in Japanese, colloquially called the "flap" consonant. For English-learning infants, some native language place-of-articulation contrasts remain difficult to discriminate even later into the first year of life. For example, English-learning 7- and 10-month-old infants fail to discriminate native English /da/ and /ɾa/ even though they have different places-of-articulation—/da/ is an alveolar articulation and /ɾa/ is dental articulation (Polka, Colantonio, & Sundara, 2001). These findings suggest that early native language learning involves a reorganization of infants' perceptual categories in the first year of life.

Kuhl and colleagues (Kuhl, Conboy, Padden, Nelson, & Pruitt, 2005) suggest that tangible individual differences in early phoneme discrimination are predictive of later language outcomes. In a correlational study, Kuhl et al. (2005) found that native and non-native phoneme discrimination at 7-months predicted infants vocabulary sizes at 18 and 24-months. At 7-months, infants were tested on their ability to discriminate native language phonemes (/ta/ vs /pa/) and non-native phonemes (Mandarin /ɕi/-/tʃi/) in two separate experimental sessions. Infants who demonstrated better native discrimination and worse non-native discrimination at 7 months had larger receptive and productive vocabularies at 18

and 24 months. This work supports the hypothesis that becoming a native language listener involves learning to ignore differences in speech sounds that are not relevant to the native language, as with the non-native contrast, and precisely categorizing the phoneme contrasts present in the infants' native language. Importantly, this phoneme categorization process has implications for later language outcomes and may be heavily reliant on infants' language environment across early development. Mahr and Edwards (2018) found that infants' lexical processing speed and language input are predictive of later language outcomes. A recent meta-analysis by Wang and colleagues (Wang, Williams, Dille, & Houston, 2020) showed a small meta-analytic effect of adult speech input on vocabulary development, and moderate effects of conversational turn-taking and child vocalizations on language outcomes. These findings suggest that the phoneme categorization and later vocabulary development might be guided by the quantity and quality of language input infants experience.

Infants' native language categorization for auditory phoneme perception has been well documented (e.g., Werker, Yeung, & Yoshida, 2012), and recent evidence suggests that a similar categorization process occurs for audiovisual speech, as well. Danielson and colleagues (Danielson, Bruderer, Kandhadai, Vatikiotis-Bateson, & Werker, 2017) found that infants at 6 and 9 months of age pay more attention to a speaker's mouth when viewing incongruent, non-native (i.e., Hindi) audiovisual speech stimuli compared to native language speech stimuli. However, at 11 months, infants in the study no longer discriminated two incongruent, non-native Hindi phonemes. The authors suggest that infants follow a similar categorization sequence in the first year of life for audiovisual phonemes, as they do with auditory-only phonemes. Further, Teinonen and colleagues (Teinonen, Aslin, Alku, & Csibra, 2008) suggest that access to visual articulatory information enhances infants' ability to form distinct phoneme categories. From these findings, Teinonen and colleagues (2008) hypothesized that redundant visual speech information might support the formation of phoneme categories. An infant's ability to form robust phoneme categories for speech perception is an important starting point that may scaffold developmental cascades throughout language learning, and audiovisual speech information may facilitate this fundamental, developmental process.

Thus, in Chapter 4 I tested the overarching hypothesis that audiovisual speech perception undergoes a process of phoneme categorization in the first year of life, and that the resulting perceptual narrowing to native language phoneme categories facilitates later vocabulary outcomes. I assessed infants' native and non-native phoneme discrimination using audiovisual speech stimuli and hypothesized that infants will have better discrimination for the native phoneme contrast as compared to the non-native phoneme contrast. This hypothesis would provide a conceptual replication of Kuhl et al. (2006), but with audiovisual stimuli. I also assessed longitudinal correlations for infants' phoneme discrimination at 7 months with their vocabulary development at 13 and 18 months, with the prediction that native language phoneme discrimination should

be positively correlated with later vocabulary measures, and that non-native phoneme discrimination will be negatively correlated with later vocabulary measures. Additionally, to further explore how audiovisual speech perception in the first year of life may influence later language outcomes, I combined the infants' McGurk Effect data from Chapter 1 with these data from Chapter 4 to measure the relationships between infants' audiovisual integration, phoneme perception, and vocabulary outcomes. I hypothesized that infants who readily integrate audiovisual speech information in early infancy will show better native language phoneme perception at 7 months, as well as greater vocabulary outcomes.

Methods

Participants

The same English-monolingual participants who participated in Chapter 1 returned to the lab at 7 months to participate in two separate speech perception tasks within one week's time. Infants participated in a native language phoneme discrimination task and non-native phoneme discrimination task (e.g., Kuhl, Stevens, Hayashi, Deguchi, Kiritani, & Iverson, 2006), both using audiovisual speech stimuli (e.g., Danielson, Bruderer, Kandhadai, Vatikiotis-Bateson, & Werker, 2017). Additionally, the parents of these infants were contacted via email when their children were 13 and 18-months-old to complete the *MacArthur Bates Communicative Development Inventory* as a follow-up measure to assess subsequent expressive vocabulary development. Power analyses of studies using a similar paradigm for auditory-only phoneme discrimination (e.g., Narayan, Werker, & Beddor, 2010; Polka, Colantonio, & Sundara, 2001) suggest that $n=24$ is sufficient for observed power of .80. I initially recruited 45 infants for Chapter 1 and Chapter 4 of this dissertation and planned to have robust statistical power for the analyses in this chapter. Due to Covid-19 restrictions, only 26 infants from Chapter 1 were able re-visit the laboratory for the current study (19 infants' appointments had to be cancelled due to laboratory shutdowns). Of these 26 infants, 16 infants contributed useable data for both the native and non-native phoneme contrasts and were thus included in the following analyses.

For inclusion in longitudinal analyses, infants must have contributed useable data for at least four out of the seven longitudinal measures which compose Chapter 1 and Chapter 4. The seven measures and the number of data points in each bin are as follows: 1) Infants' fusion looking time difference score at 4.5 months ($n=21$), 2) auditory looking time difference score at 4.5 months ($n=21$), 3) native phoneme preference score at 7 months ($n=13$), 4) non-native phoneme preference score at 7 months ($n=12$), 5) receptive vocabulary score at 13 months ($n=20$), 6) productive vocabulary score at 13 months ($n=20$), and 7) productive vocabulary score at 18 months ($n=18$). Twenty-one of the 45 infants recruited met these criteria to be eligible for inclusion in the longitudinal analyses.

Stimuli

Infants participated in two phoneme discrimination tasks: a native contrast and a non-native contrast. The native English contrast consisted of the postalveolar approximate /ɹ/ (as in the word *rice*) and the alveolar lateral approximate /l/ (as in the word *lice*). The non-native contrast for English-learning infants consisted of a Hindi contrast (i.e., /d̪/ and /d/). The /d̪/ phoneme is the unaspirated voiced, dental stop, and the /d/ phoneme is the unaspirated voiced, retroflex stop. Indeed, these two Hindi phonemes are visually distinctive for young children (Danielson et al., 2017) and are minimally contrastive in Hindi; /d̪al/ is the word for lentils and /dal/ is the word for branch. In previous work, Werker and colleagues (Werker & Tees, 1984; Werker & Lalonde, 1988) have found that English learning 6- to 8-month-old, but not 10- to 12-month-old infants, can discriminate these non-native Hindi /d̪/ and /d/ phonemes. Both contrasts for the current experiment can be discriminated based on the visual speech information, and the perception of both contrasts can be altered by native language experience (Werker & Tees, 1984; Kuhl et al., 2006). All stimuli were recorded and processed in the same manner as in Chapter 1 by a female trilingual English/Punjabi/Hindi speaker. The speaker has 12 years of Hindi language experience and is fully fluent. Moreover, the Hindi phonemes used in the current experiment (/d̪/ and /d/) are also present in Punjabi (Ohala & Ohala, 2001), which is the speaker's native language.

Apparatus

Both experimental sessions took place in a dimly-lit sound-attenuated booth and, similarly to Chapter 1, the infant participants were seated on their caregiver's lap throughout the session. Visual stimuli were presented on a centrally located monitor, and the auditory stimuli were emitted from speakers directly behind the visual display. A video camera recorded the infant's looking behavior and simultaneously relayed the infants' real-time looking behavior to the experimenter. Like in Chapter 1, the experimental stimuli were administered using Habit2 (Oakes et al., 2019), and the experimenter coded the infants' looking behavior in real-time throughout the experiment.

Procedure

Infants were randomly assigned to counterbalanced conditions of habituation stimulus, test order, and session order. I planned for infants to participate in both the native language phoneme discrimination task and the non-native phoneme discrimination task on two separate days within a week's time (Kuhl et al., 2006). Although the majority of the infants completed their sessions within the planned timeframe (average of 3.21 days between visits), due to scheduling conflicts, one infant completed both sessions on the same day with a 10-minute break between sessions and one infant completed the sessions 8 days apart. During the testing sessions, infants were habituated to an audiovisual presentation of one of the phonemes (e.g., /ɹ/) and then received two randomized

audiovisual test trials: one *same* trial with the same audiovisual phoneme (/ɹ/) and one *switch* trial with the contrasting audiovisual phoneme (/l/). The habituation criteria and parameters of the infant-controlled procedure were identical to those used in Chapter 1 and the dependent measure was each infants' looking time for the same and switch trials. In this habituation paradigm, a longer looking time to the switch trial indicates accurate discrimination of the two phonemes.

Parents of the participants from Chapter 1 and Chapter 4 were contacted via email to participate in follow-up vocabulary surveys at 13 and 18 months. Parents were emailed a link to an online version of the *MacArthur Bates Communicative Development Inventory*. These surveys were emailed to the parents and administered using HIPPA compliant Qualtrics survey software. Additionally, infants' McGurk Effect data from Chapter 1 was imputed for longitudinal comparisons.

Results

I ran paired-samples t-tests for both the native contrast (/ɹ/ vs /l/) and the non-native contrast (/d/ vs /q/) using the infants' looking times for the same and switch trials. I also ran a bivariate correlation between infants' looking time differences to measure if native phoneme discrimination was negatively correlated with non-native phoneme discrimination, as found by Kuhl and colleagues (Kuhl et al., 2005). Infants did not show evidence of perceptual discrimination for either the native phoneme contrast ($t(15)=0.439$, $p=0.667$, $d=0.110$) or the non-native phoneme contrast ($t(15)=0.496$, $p=0.627$, $d=0.124$) nor was there an effect of the infants' difference score measures between the native and non-native phoneme discrimination tasks ($t(15)=0.195$, $p=0.848$, $d=0.049$). Similarly, there was no significant negative correlation between infants' native and non-native phoneme discrimination ($r=0.218$, $p=0.806$).

To explore whether task performance is correlated with later vocabulary development (e.g., Kuhl et al., 2005), I ran bivariate correlations between infants' differences scores (same trial looking time – switch trial looking time) and their receptive vocabulary at 13 months, as well as their production vocabulary at 13 and 18 months. There were no significant bivariate correlations between infants' task performance on the native phoneme contrast and their receptive vocabulary at 13 months ($r=-0.456$, $p=0.159$), productive vocabulary at 13 months ($r=-0.176$, $p=0.604$), or productive vocabulary at 18 months ($r=-0.290$, $p=0.314$). There were also no significant bivariate correlations between infants' task performance on the non-native phoneme contrast and their receptive vocabulary at 13 months ($r=-0.423$, $p=0.195$), productive vocabulary at 13 months ($r=-0.500$, $p=0.118$), or productive vocabulary at 18 months ($r=-0.014$, $p=0.961$). Additionally, unlike the findings by Young and colleagues (Young, Merin, Rogers, & Ozonoff, 2009) there was no significant relationship between infants' McGurk fusion preference at 4.5 months and their productive vocabulary at 18 months ($r=-0.075$, $p=0.791$).

Longitudinal Analysis

To assess the relationships between infants' audiovisual speech integration, native and non-native phoneme perception, and vocabulary outcomes, I used a series of analytic approaches. First, since some of the measures in our analysis had different experimental designs and dependent variables, I standardized all dependent measures to Z-Scores. I next executed a k-means cluster analysis in JASP (JASP Team, 2020) using the Hartigan-Wong clustering algorithm on infants' auditory and fusion preference at 4.5 months, since this is the beginning point of the longitudinal analysis. The resulting Akaike Information Criterion (AIC=22.12) and Bayesian Information Criterion (BIC=30.48) for the k-means cluster analysis were lowest at a cluster size of four (*Figure 4.1 & Figure 4.2*). Lower AIC and BIC values are indicative of greater model generalizability, and model fit metrics ($R^2=0.847$) were robust.

The four clusters emerged as follows (*Figure 4.3 & Figure 4.4*). Cluster 1 (hereafter called "FamiliarPref") was characterized by a familiar auditory preference ($M=-1.453$) and a familiar fusion preference ($M=-1.390$). Cluster 2 ("AudNovel") had a strong auditory preference ($M=1.564$) and a null fusion preference ($M=0.109$). Cluster 3 ("FusionNovel") was characterized by null preferences the auditory stimuli ($M=-0.319$) and a novelty preference for fusion stimuli ($M=0.535$), and Cluster 4 ("FusionFam") had a null auditory preference ($M=0.396$) and a strong negative polarity for fusion preference ($M=-2.303$), indicating these infants treated the fusion test stimuli as familiar.

After standardizing our dependent variables and conducting the k-means cluster analysis, I assessed how audiovisual integration at 4.5 months related to infants' phoneme discrimination and vocabulary development over the first 18 months of life using growth mixture modeling in R statistical software (R Core Team, 2020) using nlme and lme4 packages (e.g., Ross-Sheehy, Reynolds, & Eschman, 2020). I first computed a growth model, with the infants' age as a fixed effect, and then I computed six separate models with Age (4.5-months, 7-months, 13-months, or 18-months), Measure (fusion preference, auditory preference, native phoneme preference, non-native phoneme preference score, receptive vocabulary at 13 months, productive vocabulary at 13 months, and productive vocabulary at 18 months), and Cluster (grouping from the k-means cluster analysis) as fixed effects and Participant as a random effect (see Table 1). I used an ANOVA for model comparison and the best fitting model was composed of Age and Measure as fixed effects and a random slope with Participant nested within AgeGroup ($\chi^2=6.814$, $p<0.01$). While this model was the best fit for our data, the model did not possess any statistically significant main effects or interaction terms (see Table 2). These results suggest that there are no significant longitudinal trends between infants' audiovisual integration, phoneme perception, and vocabulary outcomes (*Figure 4.5*).

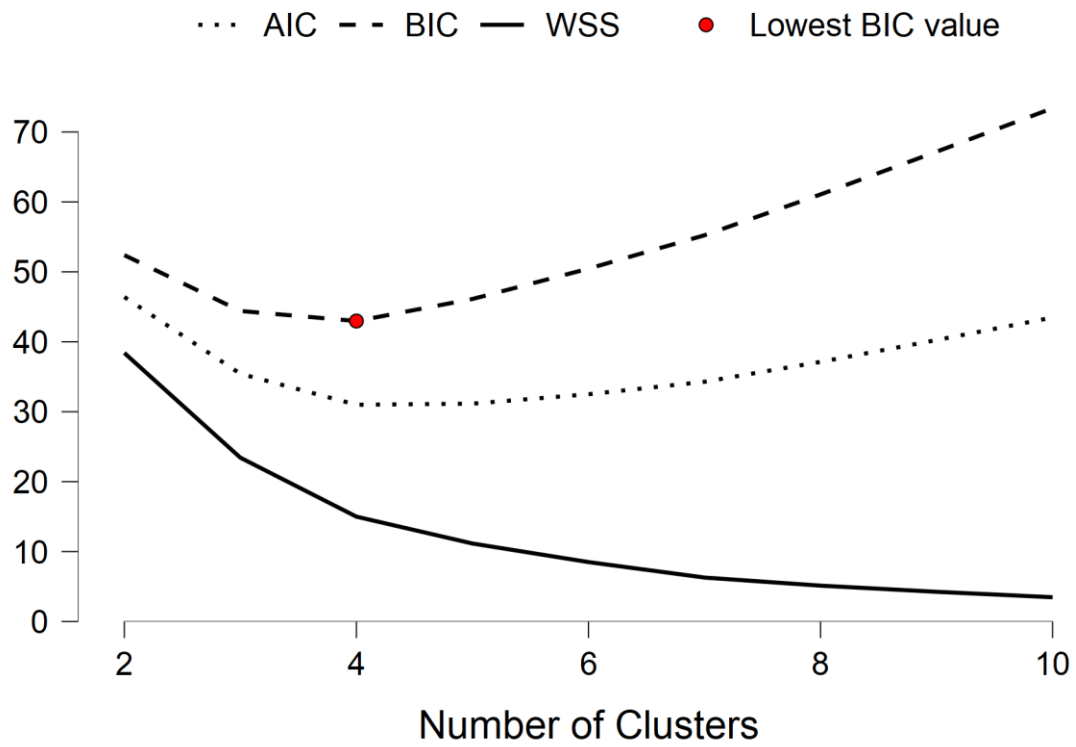


Figure 4.1. Elbow method plot for k-means cluster values derived from the lowest Bayesian Information Criterion value which suggests that four clusters are optimal in the observed data.

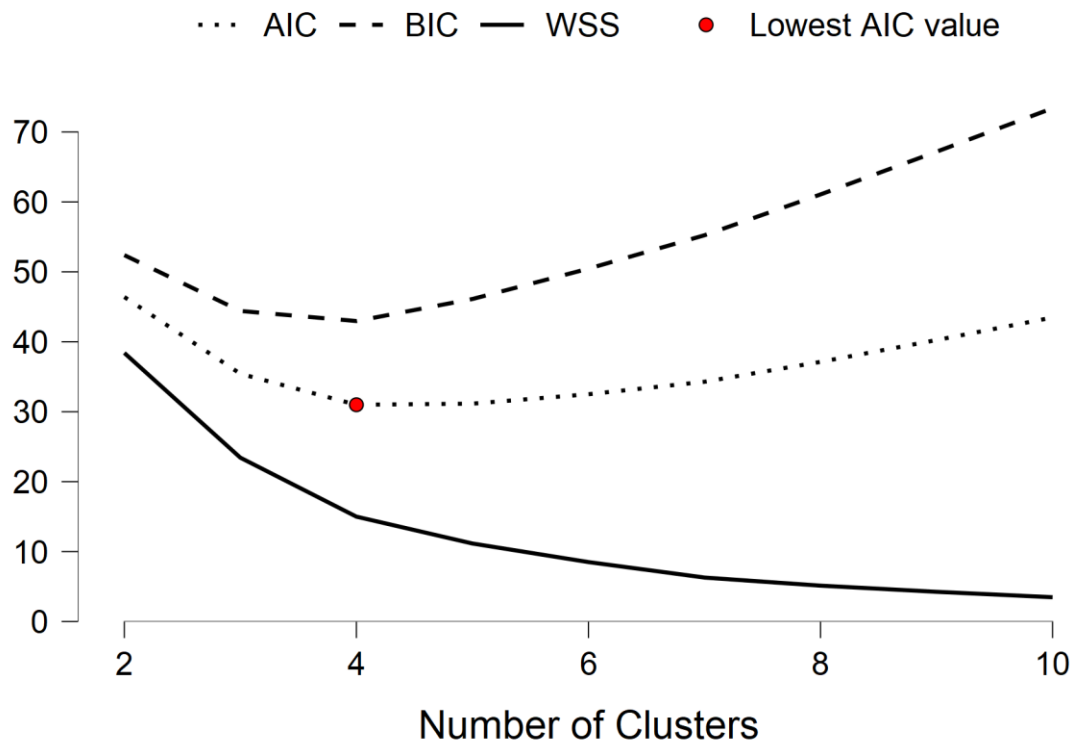


Figure 4.2. Elbow method plot for k-means cluster values derived from the lowest Akaike Information Criterion value which also suggests that four clusters are optimal in the observed data.

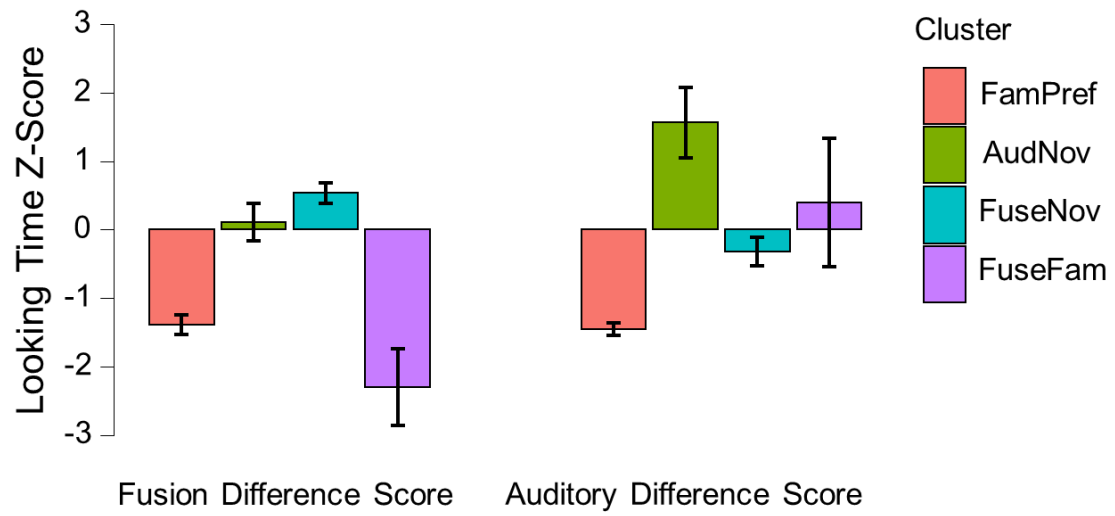


Figure 4.3. Mean and standard deviation values for the auditory and fusion preference data that compose each of the four clusters.

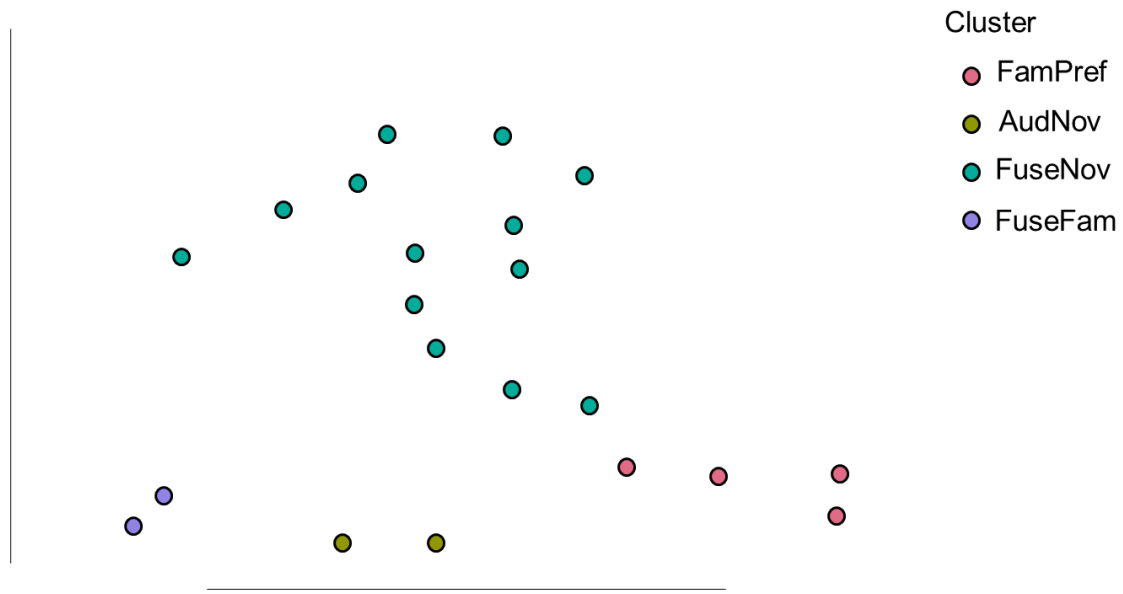


Figure 4.4. A t-Distributed Stochastic Neighbor Embedding (t-SNE) plot of the four clusters defined from our longitudinal data. Note: t-SNE reduces high-dimensional data to a 2D visual representation, thus, the specific organization shown here is conceptual.

Table 1. Model summary descriptions of all growth mixture models, with the best fitting model (m6) in italics.

Model	AIC	BIC	LogLik	p-value	Fixed Effects	Random Effects
m1	340.05	351.36	-166.02	N/A	Age	Participant
m2	343.31	360.28	-165.65	0.690	Age*Measure	Participant
m3	344.3	364.1	165.15	0.316	Age*Measure + Cluster	Participant
m4	340.94	366.39	-161.47	0.458	Age*Measure + Cluster	Participant Age
m5	347.28	375.57	-163.64	1.000	Age*Measure*Cluster	Participant
<i>m6</i>	<i>339.49</i>	<i>362.11</i>	<i>-161.74</i>	<i>0.009*</i>	<i>Age*Measure</i>	<i>Participant Age</i>
m7	343.87	377.81	-159.94	0.024*	Age*Measure*Cluster	Participant Age

Table 2. Model summary of the best fitting growth mixture model (m6).

Model Fit			
Observations	AIC	logLik	<i>p</i> -Value
125	354.91	-171.45	<0.021
Random Effects			
Group	Name	Variance	Std.Dev
Participant	(Intercept)	0.197	0.444
Fixed Effects			
	Estimate	Std.Error	t-value
(Intercept)	1.113	1.539	0.471
Age	-0.393	0.560	0.485
Measure	-0.330	0.365	0.367
Age:Measure	0.131	0.151	0.388

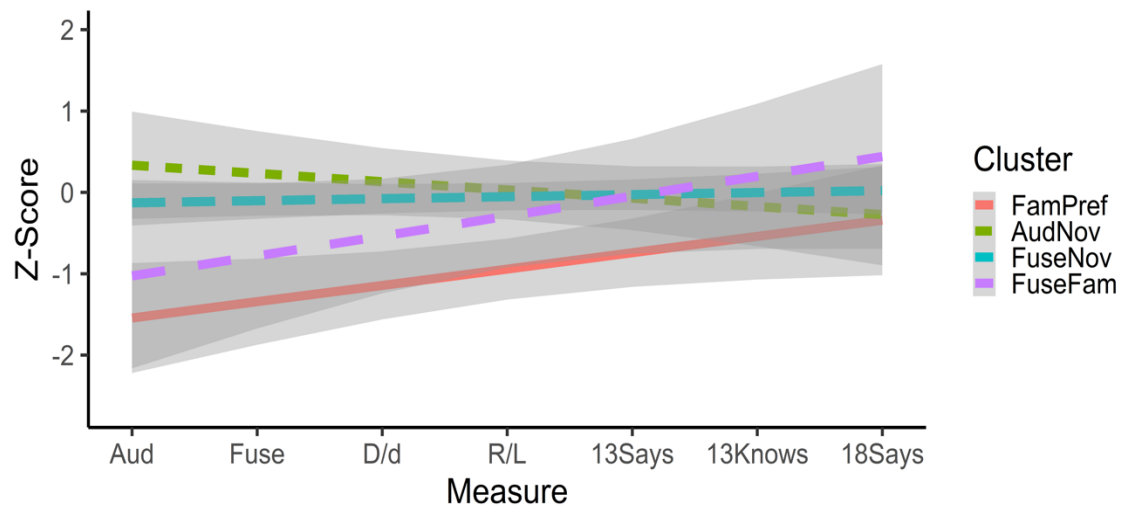


Figure 4.5. Clustered mean growth model Z-Scores indexed by Measure across the first 18 months of life. Note: Dark grey ranges indicate 95% confidence intervals.

Discussion

In Chapter 4, I first considered infants' native and non-native phoneme perception for audiovisual speech stimuli. While previous research using auditory-only phonemes (Kuhl et al., 2005) suggests that infants more readily discriminate native language phonemes as compared to non-native language phonemes, our empirical data from Chapter 4 failed to replicate this seminal finding with audiovisual speech. Further, I assessed infants' later vocabulary development as a function of their phoneme perception, and I similarly failed to replicate the previous finding that native language phoneme discrimination should be positively correlated vocabulary outcomes and non-native phoneme discrimination should be negatively correlated with later vocabulary measures (e.g., Kuhl et al., 2005). Danielson and colleagues (Danielson et al., 2017) suggested that perceptual narrowing for audiovisual phoneme perception takes place over the first year of life, however, this process may take place at a delayed onset relative to auditory-only phoneme perceptual narrowing. Our null findings from Chapter 4 may indeed support this hypothesis—the 7-month-olds in our experiment might make less robust audiovisual mappings for the stimuli I used (/ɹ/, /l/, /d/, & /q/) throughout the phoneme discrimination experiments. While we tested 7-month-olds in order to replicate Kuhl and colleagues (2005) work examining the relationship between auditory phoneme discrimination and vocabulary size, it may be the case that as infants get older, audiovisual speech information may become more robustly represented (for an example of more robust audiovisual speech representations in 9-month-olds see Danielson et al., 2017). Interestingly, much of the literature regarding infants' audiovisual integration of speech information comes from studies using stop consonants (e.g., /b/), like our stimuli in Chapter 1. However, some phoneme contrasts (e.g., English /d/ vs /ð/; Polka, Colantonio, & Sundara, 2001) have a slower developmental timescale, suggesting that while infants show robust sensitivity to some audiovisual speech, not all speech sound discrimination abilities may follow the same developmental trajectory.

In Chapter 4, I also conducted longitudinal analyses that compiled infants' task data at 4.5 months for the McGurk Effect, their phoneme perception at 7 months, and vocabulary outcomes. While the best fitting growth mixture mode for these longitudinal data was null, this approach might be a more robust way to test longitudinal predictions in the future research. For example, Ross-Sheehy and colleagues (Ross-Sheehy, Reynolds, & Eschman, 2020) have used similar methods to assess individual differences in infants' looking phenotypes, and how these looking phenotypes follow different developmental trajectories across the first year of life. While previous work in the language learning literature often correlates task performance with later vocabulary outcomes, correlational analyses are quite limited statistically, and unable to uncover causal relationships. The findings presented here are unable to substantiate previous bivariate correlations that suggest both phoneme learning (Kuhl et al., 2005) and

audiovisual integration (Young et al., 2009) in the first year of life relate to later vocabulary outcomes. While our sample was truncated due to the Covid-19 pandemic, the lack of trending bivariate relationships in our data implies there are complex relationships between perceptual narrowing with audiovisual and the implications of later language development outcomes.

CHAPTER V
14-MONTH-OLDS' LEARNING OF MINIMAL PAIR OBJECT-
LABEL ASSOCIATIONS FROM AUDIOVISUAL SPEECH

While the previous chapters of this dissertation have dealt with the influences of audiovisual speech on phoneme- and syllable-level perceptual processes, an additional area where audiovisual speech might be facilitative for language learning is during word learning. Learning to map labels to objects is a fundamental language learning challenge human infants master within their first two years of life. Previous research has aimed to measure infants' ability to map object-label pairs in laboratory settings (e.g., Stager & Werker, 1997; Horst & Samuelson, 2008) and canonical experiments by Werker and colleagues (e.g., Stager and Werker, 1997; Werker, Cohen, Lloyd, Casasola, & Stager, 1998) pioneered the use of the Switch Task experimental paradigm. The Switch Task is composed of an infant-controlled habituation phase where infants are habituated to two novel label-object pairs. Then, infants are tested on their ability to differentiate Same Trials where the original label-object pairs are maintained, from Switch Trials where Object A is paired with Label B or vice versa. Stager and Werker (1997) found that 14-month-olds readily learn object-label associations for phonologically distinct pseudowords (e.g., *Lif* and *Neem*). Crucially, 14-month-olds were unable to learn similar-sounding minimal pairs (e.g., *Bih* and *Dih*), but were able to discriminate *Bih* from *Dih* in an object-free auditory discrimination task. Werker and colleagues (Werker, Cohen, Lloyd, Casasola, & Stager, 1998) quickly replicated this interesting finding—infants readily learn dissimilar sounding words but fail to learn minimal pairs even though infants can discriminate the auditory word forms in object-free tasks. These findings suggest there is an interesting relationship between infants' object-label mapping ability and their phonological sensitivity and further launched a rich set of empirical studies to explore how infants' might use previously learned phonological representations in object-label mapping tasks.

Werker and colleagues (Werker, Fennell, Corcoran, & Stager, 2002) replicated the seminal finding that 14-month-olds failed to learn minimal pairs. However, 20-month-olds readily learned minimal pair object-label associations—17-month-olds also succeeded in the task but were not as successful as their peers who were 3 months older. Additionally, at 14 months, infants' productive vocabulary size was positively associated with a Switch preference, which is interpreted as evidence of learning in the Switch Task. In the Switch Task, infants' average looking time for *switch* and *same* test trials are statistically compared, and infants who have a longer looking time average for *switch* trials are reasoned to have noticed that the object-label pairs from habituation had been violated, thus the interpretation that infants with a *switch* preference successfully learned the initial object-label pairs. Together, these results suggest that there is a developmental trend in infants' ability to learn minimal pair object-label associations and that infants may not be able to access robust phonological representations while learning new words, even if infants can discriminate the auditory labels.

In addition to the Switch Task findings from Werker and colleagues, contemporaneous work by Swingley and Aslin (2000; 2002) used a

mispronunciation paradigm to examine children's phonological representations. Swingley and Aslin (2000) tested 18- to 23-month-old's lexical representations by comparing looking accuracy and reaction time for correctly pronounced words (e.g., *baby*) vs mispronounced words (e.g., *vaby*). Young children's looking accuracy and reaction times were affected by mispronunciations of known words, but they still reliably identified the correct referent object on trials that contained mispronunciations. These results suggest that mispronunciations *impair* but do not *inhibit* recognition of familiar words and that lexical representations for known words are robust and well-specified. These results are consistent with the findings from Werker et al. (2002) that 17- and 20-month-olds can learn minimal pairs, such that older children access robust phonological representations during speech perception. Interestingly, Swingley and Aslin (2002) found that 14-month-olds also reliably matched mispronounced words to their referent object for both close (*vaby*) and distant (*raby*) mispronunciations. While the 14-month-olds in Swingley and Aslin (2002) were able to access fine phonetic detail in their known lexical representations, the previous work from Stager and Werker (1997) and Werker et al. (2002) suggests that 14-month-olds are less able to access the fine phonetic detail during object-label mappings.

To test whether *a priori* lexical knowledge makes object-label associations easier to learn in the Switch Task, Fennell and Werker (2003) used *ball* and *doll* as the minimal pair object labels. Fennell and Werker (2003) found a robust Switch preference, suggesting that 14-month-olds can map highly familiar objects and labels in the Switch task. As a follow up, Fennell and Werker (2004) reasoned that while *a priori* knowledge of the words and objects in Fennell and Werker (2003) facilitated object-label associative learning, it was still unclear whether previously learned semantic associations or merely familiarity with the objects in the task facilitated learning. Thus, Fennell and Werker (2004) tested infants' ability to learn minimal pairs of *doll* and a pseudo-word, *goll*. Importantly, half of the infants in their sample knew the word *doll* and half did not, based on parental report on the MacArthur-Bates Communicative Development Inventory (MCDI). Both groups of infants, those who explicitly knew the object-word association for *doll* and those who did not, demonstrated a Switch preference after habituation, suggesting that mere familiarity with one of the objects, the *doll*, was sufficient to bolster infants' ability to learn known minimal pairs in the Switch task. These results suggest that familiarity with object-label pairs, even without specific language knowledge about those pairs, permits infants to access the phonetic detail needed to learn object-label associations in the Switch Task.

One potential explanation for why 14-month-olds typically fail to learn object-label pairs in the Switch Task is the Resource Limitation Hypothesis (Werker & Fennell, 2004). Werker and Fennell (2004) posited that the cognitive complexity of object-label mapping in the Switch Task may be too cognitively taxing for 14-month-olds, but not for 17- and 20-month-olds. Previous experience with the object-label pairs reduces cognitive demand in associative learning, which facilitates performance on the Switch task (Fennell & Werker, 2003; 2004).

In a further effort to measure how reduced cognitive demands may facilitate associative learning, Yoshida and colleagues (Yoshida, Fennell, Swingley, & Werker, 2009) measured infants' minimal pair object-label mapping using a habituation plus Looking-While-Listening procedure (LWL; Fernald, Pinto, Swingley, Weinberg, & McRoberts, 1998) during testing. Infants were habituated to two novel object-label pairs, but rather than a test phase with *same* and *switch* trial, the LWL procedure for testing. The LWL procedure gives infants a presentation of two objects (a correct object and a distractor object) on a monitor paired with an auditory presentation to look at one of the previously learned labels (e.g., "Where is the *dax*, can you find it?"). Researchers then measure infants' eye-gaze data to assess infants' looking response, specifically how quickly and accurately infants fixate on the correct object. The use of the LWL task, as compared to the Switch Task, is less cognitively demanding since the infants do not have to maintain one of the newly learned referent objects in their working memory throughout the test phase, the other newly learned object is also on screen as a distractor. The 14-month-olds in Yoshida and colleagues' (2009) study demonstrated they were able to learn phonetically-similar object-label associations when the cognitive demands of the task were reduced.

Taken together, the Switch Task research suggests that top-down, cognitive demands influence 14-month-old's ability to learn minimal pairs in the laboratory. Rost and McMurray (2009) suggested that bottom-up, perceptual information might also be important for object-label mapping. Rost and McMurray (2009) hypothesized that 14-month-olds may still be in the process of tuning their native phoneme categories and that variability in the speech input might augment object-label mapping, similarly to how the variability aids perceptual categorization in visual categorization (Oakes, Coppage, & Dingel, 1987) and in word segmentation (Singh, 2008). While 14-month-olds in Rost and McMurray (2009) failed to learn minimal pair object-label mappings (*buk* vs *puk*) when hearing label tokens from two different speakers, they showed evidence of learning when they were habituated to 54 unique exemplars from 18 different speakers. Rost and McMurray (2009) suggested that the use of auditory stimuli from a single speaker might hinder the infants' maintenance of their phoneme categories, while the variability in the stimuli from multiple speakers may robustly and sufficiently recruit the phoneme categories for object-label mapping. In a follow-up study, Rost and McMurray (2010) controlled for acoustic variability of voice-onset time, fundamental frequency, and amplitude of the acoustic stimuli and found that the indexical variability of the speakers who produced the speech tokens provided the relevant variability that facilitated learning, rather than small acoustic differences across productions. This evidence suggests that 14-month-olds performance in the Switch Task is not solely reliant on top-down cognitive demands, but also includes interactions with bottom-up perceptual information.

Previous research on 14-month-olds' ability to learn minimal pairs suggests that this early learning hurdle can be navigated by infants in learning

situations with reduced cognitive demands or facilitative perceptual input. In the current study, I investigated the role that audiovisually redundant speech plays in augmenting learning on minimal pair learning using the Switch Task paradigm. The extant literature on perceptual learning and the preceding chapters of this dissertation suggests that audiovisually redundant speech may be useful to guide infants' language learning. Seminal work from Bahrack and colleagues (Bahrack & Lickliter, 2000) demonstrates that perceptual learning in infancy is facilitated by intersensory redundancies. Bahrack and colleagues (Bahrack, Lickliter, Castellanos, & Vaillant-Molina, 2010) expanded the *IRH* to include a task difficulty prediction as part of their hypothesis. Infants preferentially attend to multisensory information early in development, but as they gain perceptual experience with their world, infants become increasingly good at learning from single modalities. Bahrack et al. (2010) however found that infants preferentially return their attention to multimodal stimuli in challenging learning contexts. For language learning, audiovisual speech provides the multisensory redundancies for infants as they learn language.

A recent theoretical review by Woollams (2015) suggests visual speech information is included in lexical representations, and this suggestion has been evidenced in the empirical literature. A recent study by Havy and colleagues (Havy, Foround, Fais, & Werker, 2017) aimed to explore if 18-month-olds were able to learn a new word form by solely visual speech information. While the infants were unable to learn from visual speech information alone, they were able to learn from acoustic information and then generalize to visual-only word forms at test (i.e., the visual articulations of the words learned). By 30 months of age, children were able to learn new word form associations from visual speech information alone (Havy & Zesiger, 2017). These results suggest that both auditory and visual information are stored in young children's phonological and lexical representations and motivates the current experiment by assessing how audiovisual speech information might support word learning.

The Current Experiment

The current experiment examines the utility of redundant, audiovisual speech information for object-label mapping. By using audiovisual speech stimuli in the current experiment, I explored two theoretically interesting considerations around the Switch Task and multisensory learning. First, I tested how bottom-up perceptual support might facilitate object-label mapping for minimal pairs. Secondly, this chapter is a concrete test of the task-difficulty tenet of the *IRH* as proposed by Bahrack and colleagues (Bahrack, Lickliter, Castellanos, & Vaillant-Molina, 2010), such that I predicted infants should utilize the audiovisual speech information because minimal pair object-label associative learning is challenging for 14-month-olds. Thus, the current experiment is a conceptually replication of Stager and Werker (1997) across a manipulation of visible place of articulation. In Stager and Werker (1997), they found that 14-month-olds were unable to map *bih* and *dih* to novel objects. Here, I tested infant's ability to map words (*bin* vs

din and *bin* vs *pin*) to novel objects when visual speech information was present. While *bin*, *din*, and *pin* are indeed English words, they are likely to be highly unfamiliar to 14-month-olds. In a replication of Stager & Werker (1997), Pater, Stager, and Werker's (2004) found that infants failed to learn object-label associations using combinations of these three words. The key manipulation in the current study is whether the visual speech information is distinctive or similar across the novel object-label associations. In our *visually distinctive* condition (*bin* vs *din*), /b/ and /d/ have visibly different places of articulation and in our *visually similar* condition (*bin* vs *pin*), the visual information will be articulated at the same place of articulation (i.e., bilabial). In accordance with the forth tenet of the *IRH* which theorizes that infants prioritize multimodal information in challenging learning tasks (Bahrick, Lickliter, Castellanos, & Vaillant-Molina, 2010), I hypothesized that infants will rely on the audiovisual speech stimuli in the challenging context of the Switch task and successfully demonstrate learning in the *visually distinctive* condition, where the distinctive visual speech information should be facilitative, but not the *visually similar* condition, where there is minimal difference in the redundant visual speech information. I did not include auditory-only or visual-only unimodal conditions since infants do not demonstrate evidence of learning in auditory-only tasks at 14 months (Stager & Werker, 1997) and do not show evidence of learning new object-label pairs from visual speech information until 30 months (Havy & Zesiger, 2017).

Method

Participants

Thirty-one 14-month-olds participated in the current experiment. Participants were recruited through an in-house database of families who agreed to be contacted for research studies at the University of Tennessee, Knoxville. I based our power analyses on classic studies in the minimal pair literature (Stager & Werker, 1997: n=16) and recent papers that add variability or referential support in minimal pair learning (Rost & McMurray, 2009: n=22; Fennell & Waxman, 2010: Exp. 1 n=30, Exp. 2 n=9). Our analysis suggested that n=32 (16 per condition) is sufficient for .80 observed power with an alpha level of 0.05. However, due to COVID-19 closures, our sample was truncated to 31 participants.

Stimuli

Visual Stimuli

The visual objects in our experiment consisted of two, colorful novel objects (used in Hay, Cannistraci, & Zhao, 2018). I recorded video of a female, monolingual speaker articulating each of the target words (e.g., *bin*, *din*, or *pin*) and the pre/post-stimulus pseudoword (e.g., *neem*). The speaker was recorded using a Cannon Vixia HFM40 digital camcorder, and the video stimuli were edited and sequenced using Adobe Premiere. For each trial, the audiovisual speech stimulus consisted of a woman's face on the center of the screen that

was temporally synced with the auditory stimuli. The novel objects were directly below the speaker's face on the screen, and slowly moved in a figure-8 pattern (Hay et al., 2011; Hay, Cannistraci, & Zhao, 2018). The video of the speaker articulating the target words and the novel object were spliced together using Motion software.

Auditory Stimuli

All auditory stimuli were recorded, cleaned, and normalized in Praat (Boersma & Weenink, 1996). Each novel object was labeled with a target word, dependent on condition and counterbalancing order (e.g., *bin*, *din*, or *pin*). Our speaker recorded 5 unique tokens of the object's label, and these 5 unique tokens were looped once so that infants heard up to 10 tokens across an approximately 21 second trial. The decision to use 5 unique tokens is consistent with previous work using the Switch Task (Stager & Werker, 1997; Werker, Fennell, Corcoran, & Stager, 2002; Pater, Stager, & Werker, 2004). Since the auditory stimuli had to be synced in a continuous fashion with the video stimulus recordings, there were slightly different token durations ($M=581\text{ms}$, $SD=52\text{ms}$), interstimulus intervals ($M=1,699\text{ms}$, $SD=157\text{ms}$), and trial lengths ($M=21,525\text{ms}$, $SD=100\text{ms}$) for each of the four unique trial types—though, timing differences were minimal.

Apparatus

Each experimental session was housed in a dimly lit, sound-attenuated booth. All experimental protocols were administered using Habit2 (Oakes et al., 2019) and infant participants were seated on a caregiver's lap throughout the session. The visual stimuli of the talker's face and the novel objects were presented on a centrally located television screen. The auditory stimuli were played at approximately 65 dB from speakers directly behind the visual display. A video camera hidden beneath the television screen recorded the infant's looking behavior and simultaneously relayed the infants' real-time looking behavior to the experimenter. The experimenter could not see or hear the stimuli and coded each infants' looking behavior online during the experimental sessions using the Habit2 software.

Procedure

Infants were habituated to two novel object-label pairs in the Switch Task Paradigm (Werker et al., 1998). For example, Object A was labeled *bin* and Object B was labeled *din* and these labeling conventions were counterbalanced across participants. After each infant was habituated to both object label pairs, indexed by a 50% looking time decrease from criterion (measured across a sliding 3 trial window from the first trial in the habituation phase to the last 3 trials), they moved into the testing phase. The testing phase was composed of 4 *same* test trials and 4 *switch* test trials in a pseudo-randomized order across participants. On *same* trials, the object labels matched the original contingencies

the infants learned, and I predicted this familiarity with the object-label pairing should elicit lower look durations. For *switch* trials, the incorrect label was presented with the object. If infants learned the original pairing, I predicted that they would show longer looking time because the label violates the original pairing, and this interpretation is typical of Switch Task experiments (Stager & Werker, 1997; Werker et al., 1998).

The current experiment comprised two experimental conditions using a between-subjects design: a *visually distinctive* condition and a *visually similar* condition. In the visually distinctive condition, infants were given the minimal pair words *bin* and *din* to map to novel objects. Infants in the visually similar condition were given the minimal pair words *bin* and *pin* to map to novel objects. In this condition, /b/ and /p/ are both bilabial articulations with similar visual information. Broadly, this experimental design affords the ability to test two hypotheses. The first hypothesis posits that the availability of visual speech information may facilitate infants' ability to learn the new object-label pairs in both conditions. Thus, this hypothesis predicts that infants should learn both the visually distinctive and visually similar minimal pairs. An alternative hypothesis proposes that the meaningfulness of the visual speech information might facilitate learning in the visually distinctive condition, but not in the visually similar condition. Thus, I predicted that the visual speech information may support object-label mapping in the visually distinctive condition because the /b/ and /d/ have visibly different places of articulation and the visual information would not facilitate learning in the visually similar condition, because the visual speech information is not contrastive. The hypothesized failure to learn in the visually similar condition is in line with previous research that has demonstrated that 14-month-old infants generally fail to map both contrasts (Stager & Werker, 1997). For each hypothesis and contingent sets of predictions, if speech information facilitates object-label mapping, infants should demonstrate a significantly longer looking time average on the switch trials compared to the same trials, suggesting the infants successfully learned the object-label mappings.

Results

Planned Analyses

I ran a repeated measures ANOVA to compare infants' average Same vs. Switch looking times. There was no significant interaction of Same/Switch looking time average vs experimental condition $F(1,29)=1.523$, $p=0.227$, $\eta_p^2<0.001$, nor were there main effects of Same/Switch preference ($p=0.871$) or experimental condition ($p=0.828$, *Figure 5.1*).

Post Hoc Exploratory Analyses

While I did not find an interaction or main effects when examining our Switch Task data using standard analyses for the Switch Task experimental design, an interesting correlation between infants' Same-Switch looking time

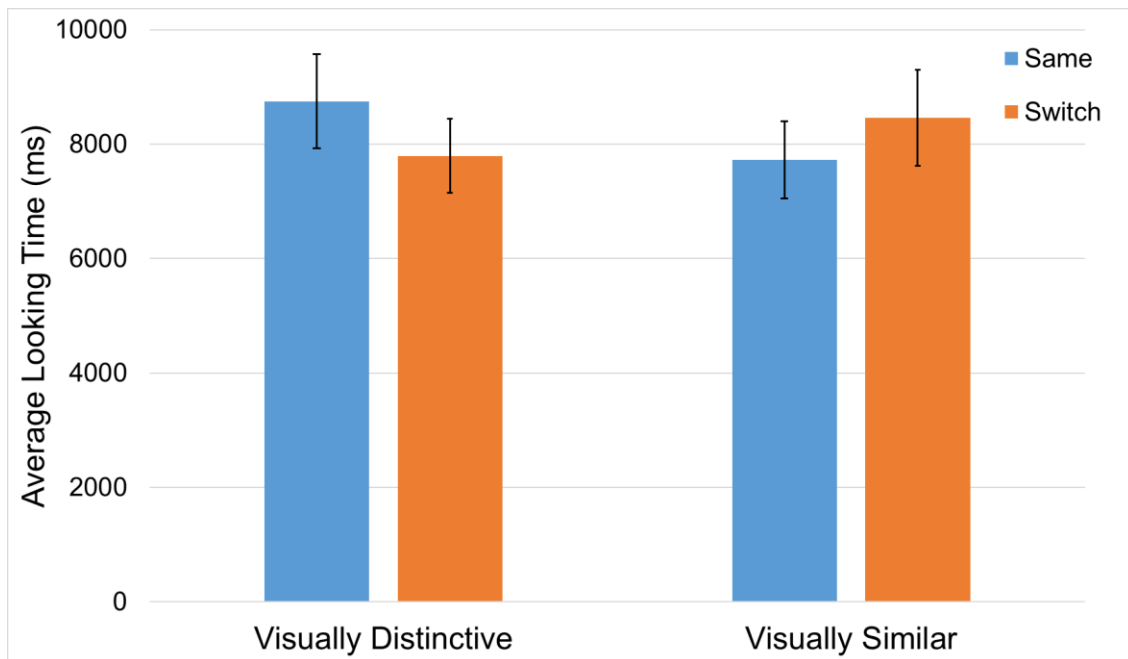


Figure 5.1. Bar graph depicting infants' looking time for Same and Switch trials across experimental conditions.

difference scores and their time to habituate emerged. Collapsed across both experimental conditions, infants were more likely to show a Switch Preference if it took them longer to habituate during the Switch task $r=-0.374$, $p=0.038$ and this trend is stable in both experimental conditions (*Figure 5.2*). This significant correlation was intriguing for two reasons. First, it suggests variability in how long it takes infants to habituate to the object-label pairings during training, and further, that longer habituation times are associated with evidence of learning. Since our task adds redundant, visual speech information to the Switch task, it may be the case that the infants are utilizing audiovisual speech during object-label mapping, and this is represented in the significant correlation.

To further explore the relationships between infants' task performance and habituation time, I conducted a post-hoc Bayesian ANCOVA. I ran an exploratory Bayesian ANCOVA because Bayesian approaches are suitable to quantify the robustness of null effects, as found in the planned analyses. I further included infants' habituation time as a covariate since infants' habituation times have emerged as a potential variable of interest in our correlation analysis. The Bayesian ANCOVA measured the relationship of Trial Type (Switch vs Same) and Condition (Visually Similar vs Visually Distinctive) as fixed effects, Participant as a random effect, and time to habituate (HabitTimeSEC) as a covariate on participants' looking time per trial. Bayesian ANCOVA works by comparing multiple models with varying effects of the predictor variables on infants' looking time, thus multiple models were analyzed (see Table 3 for model details).

Only three of the models had their model odds increased after observing the collected data—the Null Model ($m1.BF_M = 13.686$), the model including only Condition ($m2.BF_M=1.226$), and the model only including HabituationTime ($m3.BF_M=1.218$). Of these three models, the Null Model ($m1$) was most probable $P(M|data) = 0.603$. The Null Model ($m1$) was 4.90 times more likely than $m2$ (Condition; Visually Distinctive vs Visually Similar), 5.05 times more likely than $m3$ (covariate of Habituation Time), and 7.30 times more likely than $m4$ (TrialType; Switch vs Same) after observing our experimental data.

To account for model uncertainty, I performed Bayesian model averaging to test the overall effects of each predictor variable (i.e., TrialType, Condition, and Habituation Time, as well as our TrialType*Condition interaction term). Our data suggests that when averaging across the effects of all the models, the Null Model was 10.31 times more likely than models including TrialType as a predictor, 6.90 times more likely than models including Condition as a predictor, 5 times more likely than models including HabituationTime as a predictor, and 31.25 times more likely than models including the TrialType*Condition interaction term. These data suggest robust evidence for the Null Model. Critically, the covariate of interest (i.e., HabituationTime) in this exploratory analysis has no significant effect on infants' looking time during test trials (mean effect = -0.063 , 95% credible interval = $[-8350.309, -9.153]$) suggesting the previous bivariate correlation between infants' Switch preference and time to habituate may be spurious.

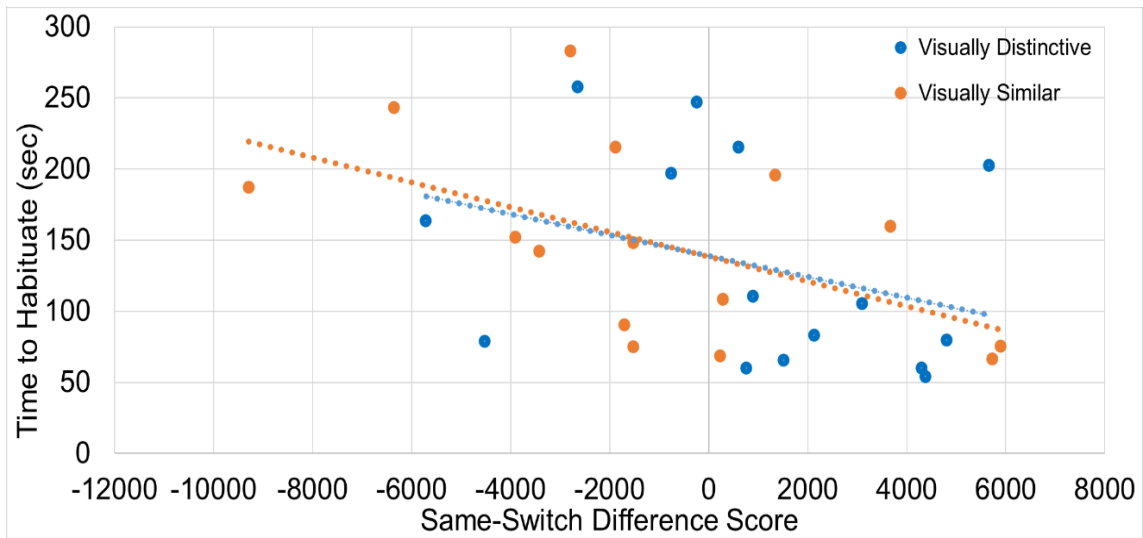


Figure 5.2. Scatterplot depicting the correlation between infants' time to habituate and their Same-Switch difference scores.

Table 3. Bayesian ANCOVA model comparison.

Models	P(M)	P(M data)	BF_M	BF₁₀	error %
m1. Null Model (incl. Participant)	0.100	0.603	13.686	1.000	
m2. Condition	0.100	0.123	1.266	0.204	0.907
m3. HabitTimeSEC	0.100	0.119	1.218	0.198	0.728
m4. TrialType	0.100	0.082	0.808	0.137	1.095
m5. Condition + HabitTimeSEC	0.100	0.027	0.247	0.044	3.177
m6. TrialType + Condition	0.100	0.018	0.161	0.029	3.634
m7. TrialType + HabitTimeSEC	0.100	0.016	0.150	0.027	2.023
m8. TrialType + Condition + TrialType * Condition	0.100	0.007	0.059	0.011	3.253
m9. TrialType + Condition + HabitTimeSEC	0.100	0.003	0.031	0.006	2.004
m10 (Omnibus Model). TrialType + Condition + HabitTimeSEC + TrialType*Condition	0.100	0.001	0.012	0.002	2.742

Note: All models include Participant as a random effect.

Discussion

Chapter 5 aimed to test the potential efficacy of audiovisual speech to facilitate 14-month-old's learning of minimal pairs. Infants were habituated to visually similar or visually distinctive minimal pairs in the Switch Task and since previous work suggests that 14-month-olds struggle to learn minimal pairs, I hypothesized that the redundant, audiovisual speech stimuli in our experiment would augment object-label mapping in the visually distinctive condition, but not in the visually similar condition. This hypothesis aimed to test a recent addition to Bahrick and colleagues' *Intersensory Redundancy Hypothesis (IRH)* (Bahrick et al., 2010), that infants should preferentially attend to multisensory information when task difficulty increases. Our mixed ANOVA found no difference of *same/switch* looking time averages between conditions, though I found a significant correlation between infants' time to habituate and their *switch* preference. However, our exploratory Bayesian ANCOVA provided robust support of the Null Model, rendering the observed bivariate correlation potentially spurious. While previous evidence suggests that audiovisual speech augments language learning in infants in a multitude of empirical queries (e.g., Teinonen et al., 2008; Lewkowicz & Hansen-Tift, 2012; Young et al., 2009; Golan et al., 2018), our null findings suggest that audiovisual speech in our task did not support minimal pair object-label mapping for 14-month-old English learning monolingual infants.

One limitation of these results is that the looking time measures used lack fine-grained information about precisely where, and for how long, infants look the speaker's face, mouth, and the referent object throughout the task. Infants who look at the speaker's mouth while the targeted pseudoword is being spoken may be able to use the visual speech information to facilitate learning more efficiently than infants who do not look at the mouth during speech events. A pattern of looking to the mouth during word articulation, coupled with looking at the referent object during silent portions of habituation would seemingly be a useful looking strategy employed by infants to learn in the experimental task. Thus, future work that capitalizes on eye-tracking or a learning phase where infants have designated times to view the visual articulatory information and the referent objects separately may be a more accurate way to investigate infants' ability to learn object-label minimal pairs from audiovisual speech.

In this experiment, I hypothesized that bottom-up influences may support infants' ability to learn object-label pairs. This hypothesis assumed that linguistic representations are amodal, and I made this assumption because infants must recruit linguistic information across sense modalities (e.g., auditory, visual, motor) for comprehension and production. Indeed, previous evidence suggests that infants can access visual speech representation during word learning and recognition (Havy & Zesiger, 2017). However, our results suggest that visual speech information may not be encoded as robustly during object-label mapping at 14 months—these audiovisual mappings may become more robust over repeated exposure that were not captured during the experimental manipulation.

CONCLUSION

The goal of this dissertation was to examine how infants perceive audiovisual speech across the first two years of life, and how these audiovisual redundancies may be used in the service of language learning. This dissertation was inspired by theoretical work, from Gibson's ecological approach to development (2000, 2003) to Bahrick and Lickliter's *Intersensory Redundancy Hypothesis* (2000, 2004), as well as previous empirical work focused on infants' language development (e.g., Burnham & Dodd, 1996, 2004; Kuhl et al., 2005; Stager & Werker, 1997). The language learning tasks selected to study infants' language development were chosen since previous research suggests these tasks capture the mechanisms of language learning challenges infants face in the first two years of life. Here, I examined the relationships between how infants perceive audiovisual speech, perceptually narrow to the important sounds in their native language, learn object-label pairs, and finally develop their receptive and productive vocabularies.

Chapter 1

In Chapter 1, I measured infants' perception of the McGurk Effect. I utilized a within-subjects design because the adult literature on the McGurk Effect suggests that some participants do not perceive the McGurk fusion, but rather perceive the auditory /ba/ stimulus present in the canonical McGurk stimuli. This experimental design allowed us to assess infants' perception of the McGurk fusion percept or the auditory /ba/ percept. Our analysis suggests that infants in our experiment perceive the auditory /ba/ stimulus when presented with the canonical McGurk stimuli. However, there were interesting patterns of data across our participants; some participants dishabituated to test stimuli that suggest they perceive the McGurk Effect during habituation, while other infants dishabituated to both test stimuli. These results suggest that individual differences in the perception of the McGurk Effect are present in early infancy, and these differences may influence infants' language learning outcomes.

Chapter 2

In Chapter 2, I followed up our empirical results from Chapter 1 with a meta-analysis to quantifying the effect size of infants' perception of the McGurk Effect. Since the results in Chapter 1 were did not follow previous findings in the literature, the meta-analysis allowed us to collate research records across published and unpublished data to get a better sense of the rate at which infants perceive the McGurk Effect. I found a moderate, cumulative meta-analytic effect for infants' perception of the McGurk Effect, and there was no evidence of publication bias or data manipulation in our analyses. I also conducted subgroup analyses, and while there were no statistically significant sources of heterogeneity in our analysis, the average effect sizes for different stimulus compositions used in McGurk Effect research were particularly interesting. The stimulus combination of the canonical McGurk Effect had the lowest cumulative

effect size, while other stimulus combinations that may elicit combination percepts or visually-dominant percepts had larger effect sizes. These results beg the question of whether different stimulus combinations that create the fusion effect ought to be treated as orthogonal from stimulus combinations that create other illusory effects.

Chapter 3

Chapter 3 follows up the results of Chapters 1 and 2 by exploring a wide range of audiovisual speech illusions as well as individual differences in the perception of the McGurk Effect using a *Dynamic Neural Field (DNF)* model. I created a neural field architecture that integrates auditory and visual speech information, and I simulated adult and infant to assess the real-time neural and perceptual processes that contribute to humans' perception of the McGurk Effect. In the adult model, I simulated a wide range of illusory effects to demonstrate how a neural architecture for audiovisual speech can support many different perceptual outcomes. I then simulated an infant model in a habituation task (as in Chapter 1) to test their perception of the McGurk Effect, and I added a working memory neural fields, as well as an autonomous looking system to simulate the interactions between memory and perception that support infant controlled habituation. Looking behavior in the autonomous looking system is based on the output of activation from the perceptual and working memory fields of the model. To explore the individual differences in infants' perception of the McGurk Effect, I created two model tunings—one that demonstrates a McGurk fusion preference and that demonstrates a preference for the auditory /ba/ stimulus. Based on these two models, I demonstrated that the relevant individual differences in McGurk perception can emerge from differences in local neural field parameter values of local excitation and lateral inhibition, as well as minute differences in recurrent connections between neural fields.

Chapter 4

Chapter 4 aims to integrate the lessons learned about individual differences in audiovisual speech perception from Chapters 1, 2, and 3, and relate these individual differences to infants' performance on native and non-native phoneme perception tasks. Within the first year of life, infants begin to home in on the relevant phoneme categories of their native language, and in Chapter 4, I tested infants' ability to discriminate native and non-native phoneme contrasts. Infants did not show evidence of discriminating the native language contrast, nor the non-native phoneme contrast, and these results fail to replicate previous work done with auditory-only phonemes (Kuhl et al., 2005). Further, I conducted a longitudinal analysis that incorporated infants' McGurk data from Chapter 1, their phoneme perception data from Chapter 4, and vocabulary outcomes at 13 and 18 months of age. I clustered infants based on their McGurk preferences in Chapter 1 to assess whether there were differences in later language outcomes based on infants' integration of audiovisual speech

information. I did not find any longitudinal relationship between McGurk preferences, phoneme discrimination and vocabulary outcomes, which also conflicts with the previous literature (Kuhl et al., 2005; Young et al., 2009).

Chapter 5

Chapter 5 continues to test the utility of audiovisual speech in the second year of life by assessing infants' ability to learn difficult minimal pair pseudowords from audiovisual speech. This empirical question was motivated by the findings from Bahrack et al. (2010), who theorized that infants preferentially attend to multisensory information during difficult learning scenarios. Thus, I examined this task-difficulty hypothesis from Bahrack et al. (2010) by testing 14-month-olds' ability to learn novel object-label pairs (e.g., Stager & Werker, 1997). In our planned analyses, I found no evidence that infants learned the novel object-label pairs from the audiovisual speech information. I did, however, find a significant correlation between infants' time to habituate to the new object-label pairs and their switch preference during the testing phase of the experiment, which indicates learning of these pairings in the chosen experimental paradigm. I then conducted a post-hoc exploratory analysis of this bivariate relationship with a Bayesian ANCOVA, and found robust support for the null Bayesian model, suggesting that the bivariate correlation between habituation time and task performance may be spurious.

Overall Conclusions and Future Directions

This dissertation aimed to substantiate and expand upon previous findings in the literature that suggest redundant, visual speech augments language learning. The empirical work presented here does not corroborate this previous evidence, and in fact, demonstrates that the influence of redundant visual speech information may be different for each individual. Further, audiovisual redundancies may be utilized by infants on a different developmental trajectory than auditory-only speech. I demonstrated individual differences in infants' perception of the McGurk Effect, but these individual differences were not related to future performance on phoneme discrimination tasks or vocabulary outcomes, suggesting that audiovisual speech does not augment language learning on a linear growth trajectory.

One interesting avenue for future research on infants' audiovisual speech perception is the interaction between infants' preferences for visual speech and the onset of canonical babbling in the first year of life. While the results of this dissertation do not support the hypothesis that audiovisual speech facilitates language learning, processing audiovisual speech may be relevant for infants' visual exploration of faces during speech. Lewkowicz and Hansen-Tift (2012) found that infants prefer to look at a speaker's mouth rather than her eyes around 8 months. Lewkowicz and Hansen-Tift (2012) reasoned that, at 8 months, infants begin to babble canonically, which is an early steppingstone to intentional language production. It may be the case that early audiovisual integration, as in

the McGurk Effect, could be related to the onset of infants preferring visual speech information, and these underlying audiovisual speech preferences might predict the onset of canonical babbling.

This hypothesized relationship between audiovisual speech preferences and canonical babbling captures the perception-action dynamics that infants must coordinate during the language learning process. In this dissertation, I focused on speech perception, but beginning to produce speech is also vital for children as they learn their native language(s). The DNF model in Chapter 3 provides a jumping off point to understand the perceptual process for audiovisual speech and the coupling of this neural architecture to a speech production architecture may provide insight as to how audiovisual speech might provide information for infants' early speech production, especially focused on the onset of canonical babbling. Indeed, DNF models are couched in perception-action couplings, and a previous DNF model created by Jackson, Spencer, and Nam (2016) has simulated differences in adult speech production between individuals who stutter and individuals with typical speech. A similar speech production neural architecture, coupled with the audiovisual speech perception architecture presented in Chapter 3 would further help untangle the complex relationships between the development of speech perception and speech production in infants. Additionally, there may be avenues to integrate speech perception and production DNF models with DNF models that offer accounts of word learning, such as the model offered by Bhat and colleagues (Bhat, Spencer, & Samuelson, 2018). The integration of speech perception and production model components with current DNF word learning models, while ambitious, could provide insight into the complex relationships between perception, action, and cognition that develop concurrently with infant language learning.

A formalization of infants' language development as offered by the DNF modeling approach would be complementary to the PRIMIR framework proposed by Werker and Curtin (2005). The PRIMIR model is a conceptual framework for Processing Rich Information from Multidimensional Interactive Representations, whereby speech is perceived by infants across a general perceptual plane, a word form plane, and a phonemic plane. Werker and Curtin (2005) reasoned that infants form multidimensional and interactive representations of linguistic information across development. For example, exemplar-like representations are processed in the general perceptual plane, and then this filtered information is transmitted to the phonemic plane. In the phonemic plane, incoming sensory information is linked to discrete representations of phonemes. This process can be reflected in the audiovisual speech processing DNF model in Chapter 3, where Hebbian learning traces are the neural representations of phonemes. The conceptual nature of the PRIMIR model synthesizes the learning processes that theoretically support a wide array of findings in the language development literature. A weakness of the PRIMIR framework is that it does not offer mechanistic accounts of how each of these levels are connected, or the processes whereby information at the different planes and filters interact. Future

work integrating the conceptual framework from PRIMIR and the neural process-based approach of the DNF language models would afford a bridge between conceptual and process-based explanations of the variability present in infants' and children's language development.

A final consideration for the body of work presented in this dissertation is that while I found null effects for the language development outcomes in typically developing participants, audiovisual speech may have clinical importance for infants who are not typically developing. D'Souza and colleagues (D'Souza, D'Souza, Johnson, & Karmiloff-Smith, 2016), found that when compared to age-matched, typically developing controls, infants with Williams syndrome were less susceptible to perceiving the McGurk Effect. D'souza and colleagues also found that infants with Down syndrome and fragile X syndrome showed general impairments in audiovisual speech processing and failed to notice when auditory and visual information were out of sync. Typically developing infants readily showed evidence of discriminating temporally synced vs non-synced audiovisual stimuli (Lewkowicz, 2010). Similarly, Bahrick and Todd (2012) suggest that children with autism spectrum disorders (ASD) show impairments in intersensory perception and that these impairments may negatively impact language outcomes. Interestingly, late onset canonical babbling is also often a hallmark of developmental disorders (Lang et al., 2019). The literature regarding sensory processing and canonical babbling similarly suggests that there may indeed be interactions between infants' multisensory processing, canonical babbling, and language outcomes. In future work, I aim to leverage the null and individual differences findings from this dissertation to study the relationships between speech perception, multisensory processes, and language development for typically and non-typically developing populations. My colleagues and I (Cannistraci, Dal Ben, Karaman, Parvanezadeh Esfahani, & Hay, 2019) have previously suggested that studying individual differences in perceptual learning and language learning may be a fruitful endeavor for understanding learning trajectories in both typical and atypical language development. The results presented in this dissertation further demonstrate the importance of assessing individual differences in language learning, especially when considering the multisensory perceptual learning processes that scaffold language learning. A convergent-methods, empirical approach as demonstrated in this dissertation that focuses on understanding individual differences in language development may prove especially advantageous for developing interventions for infants with intellectual or developmental disabilities.

REFERENCES

- Alsius, A., Paré, M., & Munhall, K. G. (2018). Forty years after Hearing lips and seeing voices: the McGurk effect revisited. *Multisensory Research*, 31(1-2), 111-144.
- Amari, S. I. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27(2), 77-87.
- Bahrack, L. E., Flom, R., & Lickliter, R. (2002). Intersensory redundancy facilitates discrimination of tempo in 3-month-old infants. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, 41(4), 352-363.
- Bahrack, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology*, 36(2), 190.
- Bahrack, L. E., & Lickliter, R. (2002). Intersensory redundancy guides early perceptual and cognitive development. *Advances in Child Development and Behavior*, 30, 153-189.
- Bahrack, L. E., & Lickliter, R. (2004). Infants' perception of rhythm and tempo in unimodal and multimodal stimulation: A developmental test of the intersensory redundancy hypothesis. *Cognitive, Affective, & Behavioral Neuroscience*, 4(2), 137-147.
- Bahrack, L. E., Lickliter, R., Castellanos, I., & Vaillant-Molina, M. (2010). Increasing task difficulty enhances effects of intersensory redundancy: Testing a new prediction of the intersensory redundancy hypothesis. *Developmental Science*, 13(5), 731-737.
- Bahrack LE, Todd JT. Multisensory processing in Autism Spectrum Disorders: Intersensory processing disturbance as a basis for atypical development. In: Stein BE, editor. *The New Handbook of Multisensory Processing*. Cambridge, MA: MIT; 2012. pp. 657–674.
- Bahrack, L. E., Todd, J. T., & Soska, K. C. (2018). The Multisensory Attention Assessment Protocol (MAAP): Characterizing individual differences in multisensory attention skills in infants and children and relations with language and cognition. *Developmental Psychology*, 54(12), 2207.
- Bergelson, E., & Swingle, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109(9), 3253-3258.
- Best, C., & Jones, C. (1998). Stimulus-alternation preference procedure to test infant speech discrimination. *Infant Behavior and Development*, (21), 295.
- Bhat, A., Spencer, J. P., & Samuelson, L. K. (2018). A dynamic neural field model of memory, attention and cross-situational word learning. In *CogSci Proceedings*.
- Boersma, P., & Weenink, D. (1996). Praat: A system for doing phonetics by computer [Computer program] (Rep. No. 132). Amsterdam, the Netherlands: University of Amsterdam, Institute of Phonetic Sciences.

- Brown, V. A., Hedayati, M., Zanger, A., Mayn, S., Ray, L., Dillman-Hasso, N., & Strand, J. F. (2018). What accounts for individual differences in susceptibility to the McGurk effect?. *PloS one*, 13(11), e0207160.
- Burnham, D., & Dodd, B. (1996). Auditory-visual speech perception as a direct process: The McGurk effect in infants and across languages. *Speechreading by Humans and Machines* (103-114). Springer Berlin Heidelberg.
- Burnham, D., & Dodd, B. (2004). Auditory–visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology*, 45(4), 204-220.
- Buss, A. T., Magnotta, V. A., Penny, W., Schöner, G., Huppert, T. J., & Spencer, J. P. (2021). How do neural processes give rise to cognition? Simultaneously predicting brain and behavior with a dynamic model of visual working memory. *Psychological Review*, 128(2), 362.
- Buss, A. T., & Spencer, J. P. (2014). The emergent executive: A dynamic field theory of the development of executive function. *Monographs of the Society for Research in Child Development*, 79(2), vii.
- Buss, A. T., & Spencer, J. P. (2018). Changes in frontal and posterior cortical activity underlie the early emergence of executive function. *Developmental Science*, 21(4), e12602.
- Buss, A. T., Wifall, T., Hazeltine, E., & Spencer, J. P. (2014). Integrating the Behavioral and Neural Dynamics of Response Selection in a Dual-task Paradigm: A Dynamic Neural Field Model of Dux et al. *Journal of Cognitive Neuroscience*, 26(2), 334-351.
- Byers-Heinlein, K., & Fennell, C. T. (2014). Perceptual narrowing in the context of increased variation: Insights from bilingual infants. *Developmental Psychobiology*, 56(2), 274-291.
- Cannistraci, R. A., Dal Ben, R., Karaman, F., Parvanezadeh Esfahani, S. P., & Hay, J. F. (2019). Statistical Learning Approaches to Studying Language Development. In *International Handbook of Language Acquisition* (pp. 51-75). Routledge.
- Cannistraci, R. A., Hay, J. F., & Buss, A. T. (2018, June). *A Dynamic Field Theory account of audiovisual speech perception: Explaining the McGurk Effect and perception of incongruous audiovisual stimuli*. Poster presented at Expanding the Field 2018: Multi-disciplinary Developmental Dynamics, Norwich, UK.
- Cannistraci, R. A., Hay, J. F., & Buss, A. T. (2019, July). *A dynamic neural field model of the McGurk Effect and Incongruous audiovisual speech stimuli*. Poster presented at CogSci 2019, Montreal, CA.
- Cutting, J. E. (1976). Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. *Psychological Review*, 83(2), 114.
- Danielson, D. K., Bruderer, A. G., Kandhadai, P., Vatikiotis-Bateson, E., & Werker, J. F. (2017). The organization and reorganization of audiovisual

- speech perception in the first year of life. *Cognitive Development*, 42, 37-48.
- Deco, G., & Rolls, E. T. (2004). A neurodynamical cortical model of visual attention and invariant object recognition. *Vision research*, 44(6), 621-642.
- Desjardins, R. N., & Werker, J. F. (2004). Is the integration of heard and seen speech mandatory for infants? *Developmental Psychobiology*, 45(4), 187-203.
- D'Souza, D., D'Souza, H., Johnson, M. H., & Karmiloff-Smith, A. (2016). Audio-visual speech perception in infants and toddlers with Down syndrome, fragile X syndrome, and Williams syndrome. *Infant Behavior and Development*, 44, 249-262.
- Eimas, P. D. (1974). Auditory and linguistic processing of cues for place of articulation by infants. *Perception & Psychophysics*, 16(3), 513-521.
- Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, 171(3968), 303-306.
- Erlhagen, W., & Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychological Review*, 109(3), 545.
- Fennell, C. T., & Waxman, S. R. (2010). What paradox? Referential cues allow for infant use of phonetic detail in word learning. *Child Development*, 81(5), 1376-1383.
- Fennell, C. T., & Werker, J. F. (2004). Infant attention to phonetic detail: Knowledge and familiarity effects. In *Proceedings of the 28th annual Boston University conference on language development* (Vol. 1, pp. 165-176). Somerville, MA: Cascadilla Press.
- Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science*, 233(4771), 1416-1419.
- Getz, L. M., & Toscano, J. C. (2021). Rethinking the McGurk effect as a perceptual illusion. *Attention, Perception, & Psychophysics*, 1-16.
- Gibson, E. J. (2000). Perceptual learning in development: Some basic concepts. *Ecological Psychology*, 12(4), 295-302.
- Gibson, E. J. (2003). The world is so full of a number of things: On specification and perceptual learning. *Ecological Psychology*, 15(4), 283-287.
- Golan, J., Daubney, K., Ballieux, H., Tomalski, P., George, R., Vitkovitch, M., Moore, D., Kushnerenko, E. (2018, July). *Predicting language development from speech perception in early childhood*. Poster presented at the XXI Biennial Conference for the International Congress on Infant Studies, Philadelphia, PA.
- Havy, M., Foroud, A., Fais, L., & Werker, J. F. (2017). The role of auditory and visual speech in word learning at 18 months and in adulthood. *Child Development*, 88(6), 2043-2059.
- Havy, M., & Zesiger, P. (2017). Learning Spoken Words via the Ears and Eyes: Evidence from 30-Month-Old Children. *Frontiers in Psychology*, 8, 2122.
- James, W. (1890). *The principles of psychology* (Vol. II). New York: Holt.
- JASP Team (2020). JASP (Version 0.14.1)[Computer software].

- Jackson, ES, Spencer, JP, & Nam, H (2016). Dynamic field-based task dynamic application (DFTaDA): An embodied model of speech production. Poster to be presented at LabPhon15, Ithaca, NY.
- Kuhl, P. K. (1991). Human adults and human infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, 50(2), 93-107.
- Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1493), 979-1000.
- Kuhl, P. K., Conboy, B. T., Padden, D., Nelson, T., & Pruitt, J. (2005). Early speech perception and later language development: Implications for the “critical period”. *Language learning and development*, 1(3-4), 237-264.
- Kuhl, P. K., & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behavior and Development*, 7(3), 361-381.
- Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, 9(2), F13-F21.
- Lang, S., Bartl-Pokorny, K. D., Pokorny, F. B., Garrido, D., Mani, N., Fox-Boyer, A. V., ... & Marschik, P. B. (2019). Canonical babbling: A marker for earlier identification of late detected developmental disorders?. *Current Developmental Disorders Reports*, 6(3), 111-118.
- Lewkowicz, D. (2010). Infant perception of audio-visual speech synchrony. *Developmental Psychology*, 46(1), 66-77.
- Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences*, 109(5), 1431-1436.
- MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics*, 24(3), 253-257.
- Magnotti, J. F., & Beauchamp, M. S. (2017). A causal inference model explains perception of the McGurk effect and other incongruent audiovisual speech. *PLoS Computational Biology*, 13(2), e1005229.
- Mahr, T., & Edwards, J. (2018). Using language input and lexical processing to predict vocabulary size. *Developmental Science*, 21(6), e12685.
- Mallick, D. B., Magnotti, J. F., & Beauchamp, M. S. (2015). Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin & Review*, 22(5), 1299-1307.
- Massaro, D. W. (1987). Categorical partition: A fuzzy-logical model of categorization behavior. In S. Harnad (Ed.), *Categorical Perception: The Groundwork of Cognition* (pp. 254-283). New York, NY, US: Cambridge University Press.

- Maurer, D., & Werker, J. F. (2014). Perceptual narrowing during infancy: A comparison of language and faces. *Developmental Psychobiology*, 56(2), 154-178.
- Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3), B101-B111.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746-748.
- McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science*, 12(3), 369-378.
- Mullen, E. M. (1995). *Mullen Scales of Early Learning* (pp. 58-64). Circle Pines, MN: AGS.
- Narayan, C. R., Werker, J. F., & Beddor, P. S. (2010). The interaction between acoustic salience and language experience in developmental speech perception: Evidence from nasal place discrimination. *Developmental Science*, 13(3), 407-420.
- Oakes, L. M., Sperka, D., DeBolt, M. C., & Cantrell, L. M. (2019). Habit2: A stand-alone software solution for presenting stimuli and recording infant looking times in order to study infant development. *Behavior Research Methods*, 1-10.
- Ohala, M., & Ohala, J. (2001). Acoustic VC transitions correlate with degree of perceptual confusion of place contrast in Hindi. *Travaux du cercle Linguistique de Copenhague*, 31, 265-284.
- Pater, J., Stager, C., & Werker, J. (2004). The perceptual acquisition of phonological contrasts. *Language*, 384-402.
- Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science*, 6(2), 191-196.
- Perone, S., & Spencer, J. P. (2013). Autonomy in action: linking the act of looking to memory formation in infancy via dynamic neural fields. *Cognitive Science*, 37(1), 1-60.
- Polka, L., Colantonio, C., & Sundara, M. (2001). A cross-language comparison of /d/-/ð/ perception: evidence for a new developmental pattern. *The Journal of the Acoustical Society of America*, 109(5), 2190-2201.
- Proverbio, A. M., Massetti, G., Rizzi, E., & Zani, A. (2016). Skilled musicians are not subject to the McGurk effect. *Scientific Reports*, 6(1), 1-11.
- R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/>
- Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science*, 17(6), 405-409.
- Rosenblum, L. D., & Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 22(2), 318.

- Rosenblum, L. D., Schmuckler M.A., and Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics* 59(3), 347-357.
- Ross-Sheehy, S., Reynolds, E., & Eschman, B. (2020). Evidence for Attentional Phenotypes in Infancy and Their Role in Visual Cognitive Performance. *Brain Sciences*, 10(9), 605.
- Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental science*, 12(2), 339-349.
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294), 1926-1928.
- Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology*, 69.
- Schöner, G., & Spencer, J. (2016). *Dynamic thinking: A primer on dynamic field theory*. Oxford University Press.
- Shaw, K. E., & Bortfeld, H. (2015). Sources of confusion in infant audiovisual speech perception research. *Frontiers in Psychology*, 6, 1844.
- Singh, L., Fu, C. S., Tay, Z. W., & Golinkoff, R. M. (2018). Novel word learning in bilingual and monolingual infants: evidence for a bilingual advantage. *Child Development*, 89(3), e183-e198.
- Singh, L., Poh, F. L., & Fu, C. S. (2016). Limits on monolingualism? A comparison of monolingual and bilingual infants' abilities to integrate lexical tone in novel word learning. *Frontiers in Psychology*, 7, 667.
- Stager, C.L. & Werker, C.F. (1997). Infants listen for more phonetic detail in speech perception than in word learning tasks. *Nature*, 388, 381-382.
- Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, 26(2), 212-215.
- Swingle, D. & Aslin, R.N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, 76, 147 – 166.
- Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition*, 108(3), 850-855.
- Tyler, M. D., Best, C. T., Goldstein, L. M., & Antoniou, M. (2014). Investigating the role of articulatory organs and perceptual assimilation in infants' discrimination of native and non-native fricative place contrasts. *Developmental Psychobiology*, 56(2), 210-227.
- Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, 104(33), 13273-13278.
- Wang, Y., Williams, R., Dilley, L., & Houston, D. M. (2020). A meta-analysis of the predictability of LENA™ automated measures for child language development. *Developmental Review*, 57, 100921.
- Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Gallés, N., & Werker, J. F. (2007). Visual language discrimination in infancy. *Science*, 316(5828), 1159-1159.

- Werker, J. F., Cohen, L. B., Lloyd, V. L., Casasola, M., & Stager, C. L. (1998). Acquisition of word-object associations by 14-month-old infants. *Developmental psychology*, 34(6), 1289.
- Werker, J. F., Fennell, C. T., Corcoran, K. M., & Stager, C. L. (2002). Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. *Infancy*, 3(1), 1-30.
- Werker, J. F., & Lalonde, C. E. (1988). Cross-language speech perception: Initial capabilities and developmental change. *Developmental Psychology*, 24(5), 672.
- Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7(1), 49-63.
- Werker, J. F., Yeung, H. H., & Yoshida, K. A. (2012). How do infants become experts at native-speech perception?. *Current Directions in Psychological Science*, 21(4), 221-226.
- Wijeakumar, S., Ambrose, J. P., Spencer, J. P., & Curtu, R. (2017). Model-based functional neuroimaging using dynamic neural fields: An integrative cognitive neuroscience approach. *Journal of Mathematical Psychology*, 76, 212-235.
- Woollams, A. M. (2015). Lexical is as lexical does: computational approaches to lexical representation. *Language, Cognition and Neuroscience*, 30(4), 395-408.
- Young, G. S., Merin, N., Rogers, S. J., & Ozonoff, S. (2009). Gaze behavior and affect at 6 months: predicting clinical outcomes and language development in typically developing infants and infants at risk for autism. *Developmental Science*, 12(5), 798-814.

APPENDIX

A more thorough treatment and explanation of the equations used in dynamic neural field models can be found in the seminal work by the Schöner and Spencer (2016) and The DFT Research Group. The explanation of the equations used in Chapter 3 of this dissertation are modeled on the Supporting Information give by Buss and Spencer (2018) and the Appendix of Buss and Spencer (2014)

Model Equations

The basic formula for a 1-dimensional neural field tuned for working memory peak formation and maintenance is given in Equation 1. The rate of activation changes in the cortical field (w) changes over time (t) at each location in the neural field (x). In the current model, the w field is a working memory field, which I will call WM for ease of explanation.

$$\tau_w \dot{w}(x, t) = -w(x, t) + h_w + S(x, t) + \underbrace{\int c_{ww}(x - x') \Lambda(w(x', t)) dx'}_{\text{dotted line}} + \underbrace{\int c_{wv}(x - x') \Lambda(v(x', t)) dx'}_{\text{solid line}} \quad 1$$

The first part of Equation 1, underlined with a dotted line, includes the neural resting level ($h_w < 0$) and stimulus input ($S(x, t)$), as well as a stabilization term ($-w(x, t)$). The resting level indicates how far a field is from the activation threshold. The stabilizing term maintains activation around an attractor state. As a system is pushed from its stable state, the rate of change goes in the opposite direction of this perturbation. The tau parameter, τ_w , captures the timescale when activation levels approach the attractor state.

The remainder of the equation specifies the excitatory (dashed line) and inhibitory (solid line) neural interactions. Neural interactions within a field are determined by the convolution of a sigmoid threshold function and a Gaussian projection. The term $\Lambda(w(x', t))$ is the convolved sigmoid value of activation at each location in field (w) and is used as the field's self-excitatory projection, while $\Lambda(v(x', t))$ is the convolved sigmoid value of activation at each point in the inhibitory field (v), used for the inhibitory projection to field (w). The sigmoid function is given by Equation 2.

$$\Lambda(w) = \frac{1}{1 + \exp[-\beta w]} \quad 2$$

In the sigmoid function, the beta (β) parameter defines the slope that transforms field activation into neural activity. The activation threshold is represented by the point where the convolved sigmoid output reaches 0.5, which is 0. With a large β value, the slope is steeper and there is a more abrupt transition in the sigmoided output. Weak levels of activation contribute relatively little to activation and peak formation, while strong levels of activation in the fields robustly interact with associated neurons.

The spread and strength of neural interactions is determined by a Gaussian interaction kernel, defined in Equation 3.

$$c(x - x') = c \exp \left[\frac{(x-x')^2}{2\sigma^2} \right] \quad 3$$

The parameter denoted by c scales the strength of the projection while the width of the interaction kernel is given by σ . Finally, these terms are integrated to factor the contributions at each location, x , from all other locations, x' (see Equation 1).

The equation for the inhibitory layer takes the same general form as the equation for the WM field:

$$\tau_v \dot{v}(x, t) = -v(x, t) + h_v + \int c_{vw}(x - x') \Lambda(w(x', t)) dx' \quad 4$$

The timescale of activation in this field is given by τ_v while its resting level is denoted h_v . In this equation, input to the inhibitory layer is the integration of above-threshold activation within the WM field, $\Lambda(w(x, t))$, with the spread and strength of this projection given by the Gaussian interaction kernel, $c_{vw}(x - x')$.

Further contributions to the dynamics of the WM field come from a Hebbian layer (HL), now added in Equation 5.

$$\begin{aligned} t_w \dot{w}(x, t) = & -w(x, t) + h_w + S(x, t) + \int c_{ww}(x - x') \Lambda(w(x, t)) dx' \\ & + \int c_{wv}(x - x') \Lambda(v(x, t)) dx' + \int c_{HL}(x - x') w_{HL}(x, t) dx' \end{aligned} \quad 5$$

The Gaussian interaction kernel, $c_{HL}(x - x')$, determines the strength and width of the projection from the HL into the WM field. The dynamics of the HL (w_{HL} ; Equation 6.0) are divided into two components (Equations 6.1 and 6.2) that capture the build and decay dynamics of the Hebbian layer:

$$\begin{aligned}\dot{w}_{HL}(x, t) &= \dot{w}_{HLbuild}(x, t) + \dot{w}_{HLdecay}(x, t) & 6.0 \\ \tau_{build}\dot{w}_{HLbuild}(x, t) &= [-w_{HL}(x, t) + \Lambda(w(x, t))] \cdot \theta(w(x, t)) & 6.1 \\ \tau_{decay}\dot{w}_{HLdecay}(x, t) &= -w_{HL}(x, t) \cdot [1 - \theta(w(x, t))] & 6.2\end{aligned}$$

The shunting term, θ , gates activation into the HL from the WM field ($\theta = 1$ when $w(x, t) > 0$, and $\theta = 0$ otherwise). With $\theta = 1$, Equation 6.1 is engaged and drives the accumulation of activation in the Hebbian layer at sites associated with above-threshold activation in the WM field. By contrast, when $\theta = 0$, equation 6.2 is engaged and activation levels in the HL decay. Separating the build and decay mechanisms approximates accumulation and depression of synaptic change (Deco & Rolls, 2004). The build timescale (e.g., $\tau_{build} = 1200$) is shorter than the decay timescale (e.g., $\tau_{decay} = 5,500$) which makes activation in the HL build more quickly relative to the rate of decay; however, both of these processes in the Hebbian layer are significantly slower than the timescales of memory peak formation in the WM field (i.e., $\tau = 20$). As inputs are presented to the WM field and peaks of activation are built, activation accumulates slowly in the HL. This accumulated activation acts as an input to the WM field which can influence the stability properties of the WM field.

VITA

Ryan Andrew Cannistraci was born in Hershey, Pennsylvania to parents Dan and Elizabeth Cannistraci. He has three siblings: Daniel, Stephen, and Lauren. Ryan graduated from Cedar Crest High School in Lebanon, Pennsylvania in 2010 and then attended Millersville University for his Bachelor of Arts degree in Psychology in 2014 under the mentorship of Dr. Shaun Cook and Dr. Shawn Gallagher. Following his undergraduate degree, he accepted a graduate teaching assistantship at the University of Tennessee, Knoxville in the Experimental Psychology program under the mentorship of Dr. Jessica Hay. Ryan earned a Master of Arts degree in Experimental Psychology in 2017 and defended this dissertation for his Doctorate in Philosophy degree in 2021. Ryan will be a Visiting Assistant Professor of Psychology at Lafayette College in Easton, Pennsylvania for the 2021-2022 academic year.